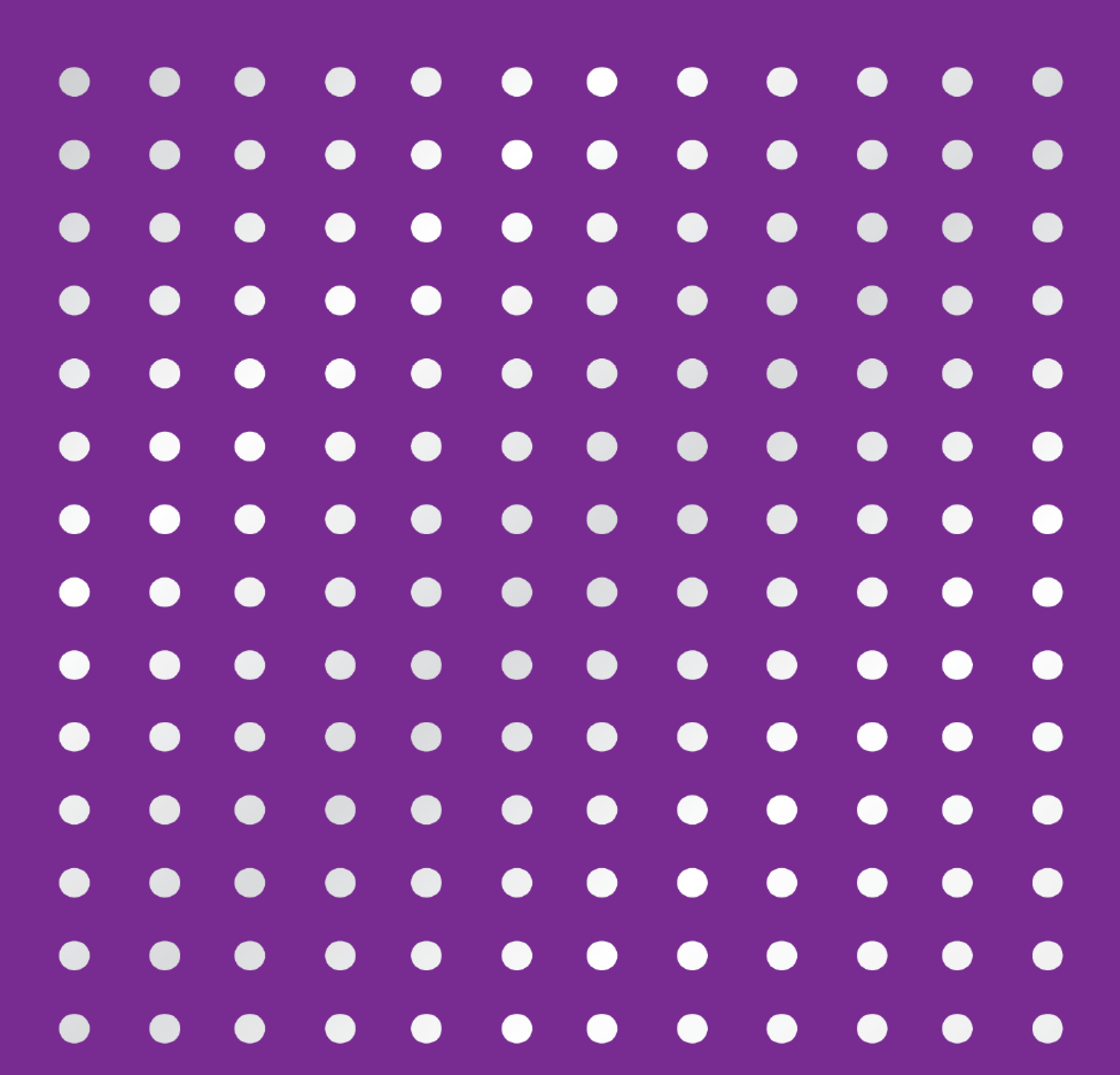
高等院校信息技术规划教材

SAS数据挖掘与分析

周　爽　贾克云　阮桂海　编著



清华大学出版社

高等院校信息技术规划教材

SAS 数据挖掘与分析

周　爽　贾克云　阮桂海　编著

清华大学出版社

北　京

内 容 简 介

本教材囊括了 SAS 编程方面极为详尽的命令语句，是数据的科学挖掘和统计分析方面的一本不可多得的教科书。

本书的前 8 章介绍了数据挖掘和统计分析所用的各类命令语句，第 9～17 章着重介绍如何用命令语句及其对话框进行常用的初高级统计和专业统计，并对统计结果进行了科学准确的分析。

本书面向全国高校统计学、医学、心理学、市场营销学、人文社会学、信息管理学及财经学等专业的本科生和研究生，可作为这些专业及其他非计算机专业学生必选的统计教材，也是数据挖掘和信息分析方面的利器。

图书在版编目（CIP）数据

SAS 数据挖掘与分析/周爽，贾克云，阮桂海编著. —北京：清华大学出版社，2008.8

（高等院校信息技术规划教材）

ISBN 978-7-302-16920-8

Ⅰ. S…　Ⅱ. ①周… ②贾… ③阮…　Ⅲ. 数据采集—统计分析—应用软件，SAS—高等学校—教材　Ⅳ. TP274

中国版本图书馆 CIP 数据核字(2008)第 009676 号

责任编辑：袁勤勇　李玮琪

责任校对：白　蕾

责任印制：何　芊

出版发行：清华大学出版社

http://www.tup.com.cn

c-service@tup.tsinghua.edu.cn

社 总 机：010-62770175

投稿咨询：010-62772015

地　　址：北京清华大学学研大厦 A 座

邮　　编：100084

邮购热线：010-62786544

客户服务：010-62776969

印 装 者：山东新华印刷厂临沂厂

经　　销：全国新华书店

开　　本：185×260　**印　张**：23.25　**字　数**：547 千字

版　　次：2008 年 8 月第 1 版　**印　次**：2008 年 8 月第 1 次印刷

印　　数：1～3000

定　　价：33.00 元

本书如存在文字不清、漏印、缺页、倒页、脱页等印装质量问题，请与清华大学出版社出版部联系调换。联系电话：010-62770177 转 3103　　产品编号：027153-01

前言 Foreword

在市场经济充满着激烈竞争的今天，统计学中的热门话题之一是数据挖掘问题。数据挖掘是一个机遇，也是一个挑战。其实，我们每位统计人员，经常在做数据挖掘工作，只是没有专门留意罢了。数据挖掘是从数据仓库中发掘那些潜在的、鲜为人知的数据规律和数理模式，其目的是在过去的经验基础上预测未来的发展趋势。例如，线性回归(linear regression)等用于预测未来，线性判别分析(linear discriminate analysis)技术用于数据分类，非线性回归技术等用于估值和抽样，从而提高市场的决策能力和成功几率。SAS(statistical analysis system)的统计方法海纳百川，其输出结果科学清晰，图形报表丰富多彩。而且SAS 9.0以上版本，比以前的纯英文SAS软件，其界面更加友好，功能更加强大，是数据挖掘和数据统计分析的锐利武器。

一旦掌握了SAS和SPSS两种知识，数据挖掘和统计分析的一切问题就能迎刃而解，社会统计学专业的学生才算长齐了双翅。为此，我们特地新编了本书、《统计分析应用教程》(ISBN: 9787302065975)和《SAS统计分析实用大全》(ISBN: 9787302064169)系列教材。前二者是"姊妹篇"，构成了普通高等教育'十一五'规划教材《社会统计分析——SAS应用教程》。这三本系列教材所介绍的知识与技术，具有很强的互补性，既体现了教材的多样性、承前启后性，又适应了不同层次、不同类型读者的需求，对课题研究及数据挖掘与分析将会奠定良好的基础。

上面所提及的三本书中的实用程序、数据和习题参考答案，可从清华大学出版社的网站下载，网址为www.tup.tsinghua.com.cn。如果难于下载可与清华大学出版社的客户服务联系，电话是010-62770175-4608。

参加编写本系列教材的成员(排名不分先后)有周爽、蔡建平、

蔡建瓴、樊爱萍、蔡建华、王肖群、蔡建琼、吴宝科、朱志海、于惠芳、崔博、顾林枝、朱秀萍、朱星萍、阮潮海、吴少宁、蔡福金、曲庆云、严康敏、孟峥、蔡楠、贾克云、杨武栓、杨靖、赵晓梅、秦小峰、智庆民、陈丹、谢力丹、阮桂海等。

作者

2008.6

目录 Contents

第 1 章 chapter 1

SAS 编程的语法知识

单击 SAS 对话框虽然可以进行统计分析，但通过用 SAS 的命令语句来编程，然后运行程序则更胜一筹。为此，本章着重介绍 SAS 的语法知识。

1.1 SAS 概 述

1. 发展史

1966 年美国 North Carolina 州立大学开始研发 SAS，1976 年成立了 SAS 研究所，1985 年 SAS 从大型机移植到计算机上。现在，SAS 和 SPSS（Statistical Package for Social Science，社会科学统计软件包）成了数据挖掘和数据统计分析的利剑。

2. SAS 的特色

（1）数据兼容性强。

（2）编程语言丰富：有 100 多种运算函数（语句），其中有：

算术运算符：＋－ * / 及 ** ；

逻辑操作符：AND，OR，NOT；

赋值语句：Z＝X＋Y；

条件语句：IF THEN/ELSE；

数组语句：ARRAY…；

循环语句：DO…END。

（3）能连读及处理多个数据集：可从各个数据集里合并观测值（OBS：个案），还能组合变量和建立子集；能处理多个输入文件。可存储会话结果和中间结果。

（4）统计方法海纳百川。

（5）强大的宏（代换）功能：大大简化了编程。

（6）计算的精度达到小数点后 11 位。

下面列举一个最简单的 SAS 程序（结构）。

程序 1.1：

```
/* 程序 1.1: */
```

```
DATA id sj;
INPUT id age v2@@ ;
  year=v2*12;
CARDS;
01 30 2500 02 29 3000 03 33 3550 04 56 3300 05 45 2800
;
PROC PRINT;
RUN;
```

运行程序 1.1 可输出图 1.1 所示的结果。



图 1.1 最简单的 SAS 程序

1.2 观测值、变量常量

1. 观测值

一份问卷、一个单一的整体、一个人、一个被测对象就是一个观测值(OBS)或称一个“个案”。每个个案是由若干变量组成。

2. 变量

一份问卷一般有几个甚至几十个问答题，一个问答题就是一个变量(Variable)。如 _id _,sex,age,income 等。

(1) 变量名：由 1～8 个有效字符组成且字母领头，后跟数字或有效的字母。但字母 @，#，$，%，^，&，* 等是无效的字符。如 sex，age，v1，location，_ab_ 等变量名是正确的。

(2) 无效的变量名：1age，1v，location1，@1，#1，%1，&2 等变量名是无效的。

(3) SAS 内部特殊的变量名：

TYPE
ERROR
_N_

(4) 变量的类型：有以下两种。

数字型：如 INPUT id sex age;

字符型：如“INPUT id sex $ age;”中的“sex $”表示性别是以 m=男性，f=女性表示。

(5) 变量赋值的特点：首次定义的长度一直有效到定义另外一种长度。见程序 1.2。

程序 1.2：

```
DATA P;             /* 将此次 DATA步处理的数据存入 WORK.p数据集(工作文件)中 */
   x='MALE';        /* 首次给变量 x赋予 4个字符的长度 */
PUT x;              /* 在 LOG窗口则输出变量 x的值为 4个字符的长度,即 MALE */
   x='FEMALE';      /* 以后,变量 x虽然输入 6个字符但只保留原来的长度 4 */
PUT x;              /* 以后,变量 x就按照原来的长度 4被记忆下来,因此此时的变量
                       x的值被截取为 FEMA */
RUN;
```

运行程序 1.2 产生图 1.2 所示的结果(见“日志”窗口)。

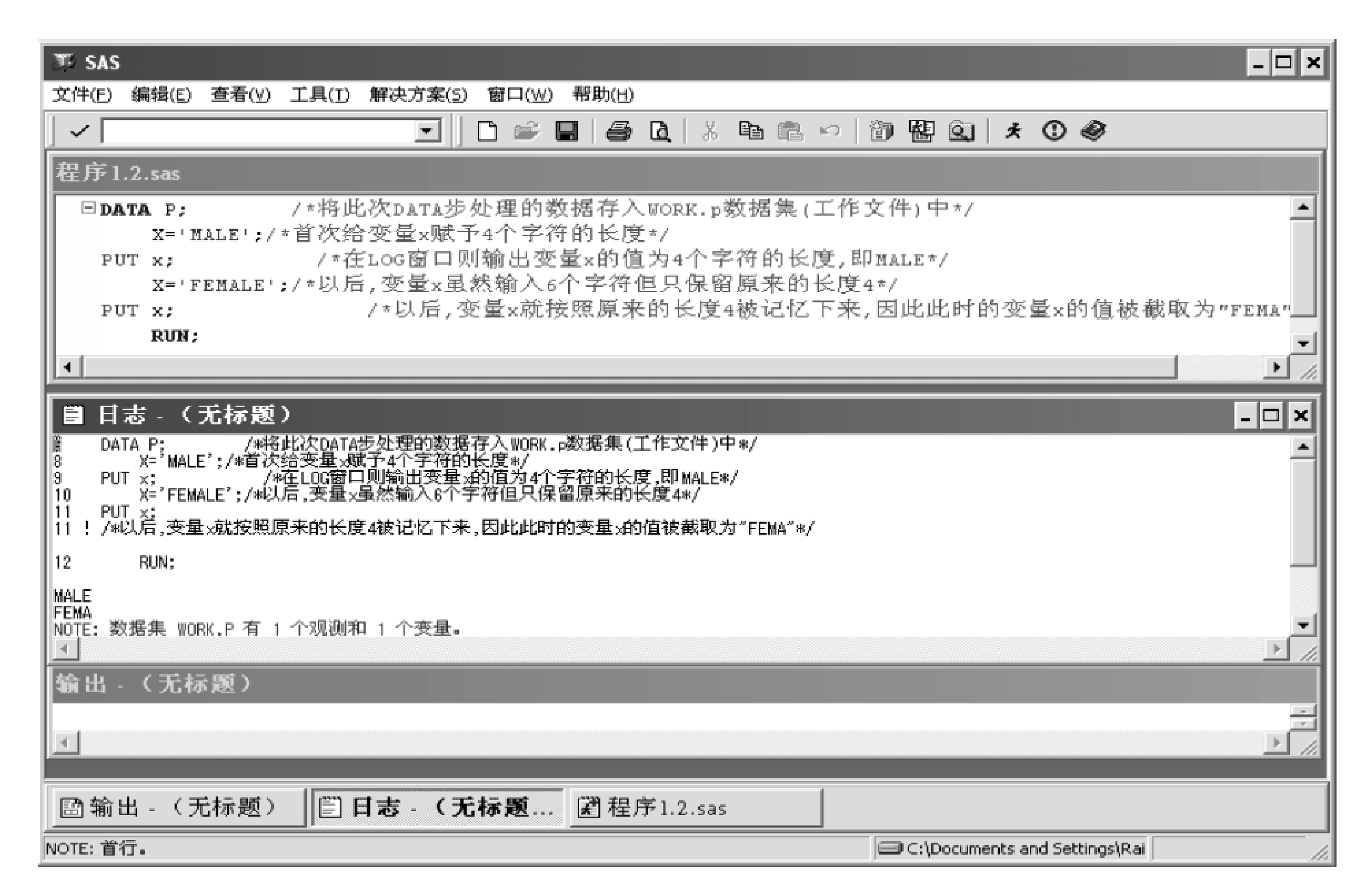


图 1.2 “日志”窗口的结果

(6) 变量的缺失值。

凡是被访对象不答或漏答的数字型变量值要输入“.”，字符型变量值要输入一个空格“ ”。见程序 1.3。

程序 1.3：

```
DATA sj2;
INPUT v1 v2 @@ ;
  year=v2 * 12;
CARDS;
01 2500 02 3000 03 2800 04 3300 05 .
;
PROC PRINT;
RUN;
```

运行程序 1.3 产生图 1.3 所示的结果(见“输出”窗口)。

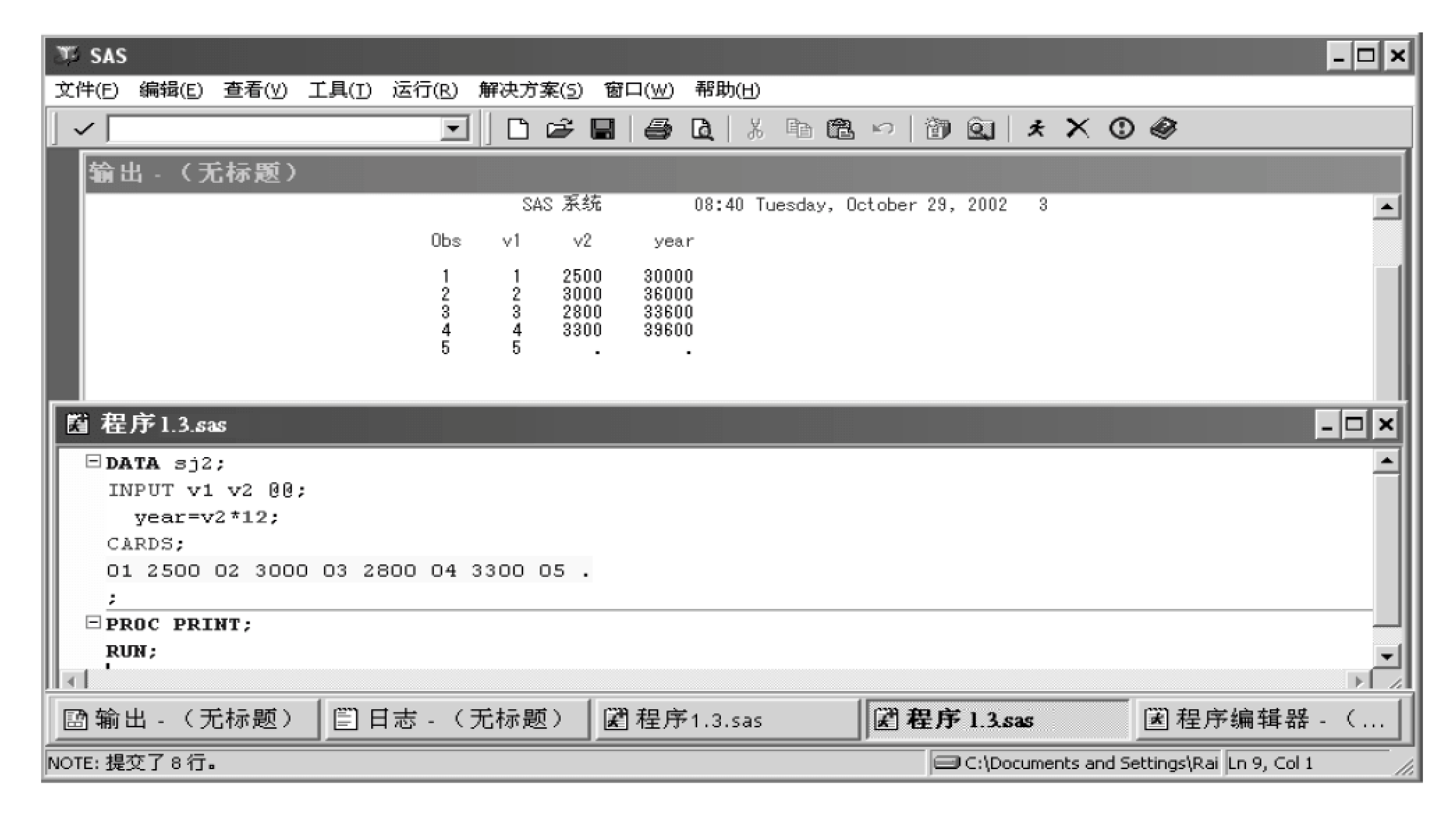


图 1.3 运行程序 1.3 产生的结果

3. 常量

常量可用于 IF、SELECT、RETURN、PUT、ERROR 等语句中，还用于赋值及求和。常量有以下 3 种语句：

(1) 数值常量：有整数、小数、科学记数法。例如 1.288E−8(相当于 $1.288 * 10^{-8}$)。

(2) 字符常量：可由 1～200 个字符组成。例如 Zhangsan，Lisiguang。

(3) 日期、时间、日期时间常量：用一对单撇号(或单引号)括起来，后跟一个日期的字母。如

```
'2006OCT10'D
'12:00'T
'12:00:18'T
'10oc2006:12:00:18'DT
```

其中，T 表示 Time(即时间)，D 表示 Date(即日期)。

1.3 SAS的操作符

操作符用于进行算术运算、比较判断、逻辑操作的指令。

1. 算术运算符

算术运算符按运算的优先顺序有：**（乘方）、*/（乘或除）、+-（加或减）。

2. 关系运算符

关系运算符有：EQ（或=）、NE（或~=）、GE（>=）、GT（或>）、LE（或<=）、LT（或<）。它们被用于关系表达式中，当表达式为真时输出1，当表达式为假时输出0。

程序1.4：

```
DATA sj3;
INPUT id month @@ ;
  year=month* 12;
  IF year< 36000 THEN C=1;ELSE c=2;
CARDS;
01 2500 02 3000 03 2800 04 3300 05.
;
PROC PRINT;
RUN;
```

运行程序1.4产生图1.4所示的结果（见“输出”窗口）。



图1.4 运行程序1.4产生的结果（见“输出”窗口）

3. 逻辑运算符

逻辑运算符也称为布尔运算符。它有以下 3 种运算符：

AND(&)、OR(|)、NOT(~)。

例如IF v1>v2 AND v2>0 THEN C+1；

　　IF v1>v2 OR v2>0 THEN C+5；

程序 1.4a：

```
DATA sj3;
INPUT v1 v2 @@ ;
 IF v1>v2 AND v2>0 THEN C1+1;            /* 如果 v1>v2 且 v2>0 则 C1+1; */
 IF v1>v2 OR v2>0 THEN C2+5;             /* 如果 v1>v2 或 v2>0 则 C2+5; */
 IF v1>v2 OR v1~=0 THEN C3+15;           /* 如果 v1>v2 或 v1 不等于 0 则 C3+15 */
 IF~ v1>v2 OR~ v1~=0 THEN C4+100;        /* 如果不是 v1>v2 或 v1 等于 0 则 C4+100 */
 z1=(3><4);                              /* 用操作符 MIN 或><取其中的最小值 */
 z2=(3<>4);                              /* 用操作符 MAX 或<>取其中的最大值 */
 CARDS;
2500 3000 2800 3300
;
PROC PRINT;
RUN;
```

运行程序 1.4a 产生图 1.5 所示的结果(见“输出”窗口)。



图 1.5　程序 1.4a 运行的结果

4. 最大值最小值的操作符

(1) 用操作符 MIN 或><取其中的最小值。例如 z1=(3><4)；

(2) 用操作符 MAX 或<>取其中的最大值。例如 z2=(3<>4)；

见程序 1.4a。

5. 数字转为字符

SAS 会自动将数字转为字符。见程序 1.4b。

程序 1.4b：

```
DATA sj4;
v1='4';
v2='6';
  Y=v1+v2;
CARDS;
PROC PRINT;
VAR Y;
RUN; /* 输出为 Y=10 */
```

1.4 SAS 数据挖掘常用的语句

SAS 编程时常用的变量表示法见表 1.1，程序的核心语句见表 1.2，程序中的操作指令见表 1.3，循环控制语句见表 1.4。

表 1.1 变量表示法

变　量　表	缩　　写	含　　义
v1 v2 v3 v4	v1—v4	从 v1 至 v4 四个变量
v1 1—2 v2 3—4 v3 5—6	(v1—v3)(3*2.)	(3*2.)表示共有三个变量，每个变量值有两位数据
v1 1—2 v2 3—4 v3 5—6	(v1—v3)(2. 2. 2.)	表示共有三个变量，每个变量值有两位数据
SAS 特殊名称	_NUMERIC_ _CHARACTER_ _ALL_	所有数字型变量 所有字符型变量 所有变量
XPA *	X-A X-NUMERIC-A X-CHARACTER-A	从 x 到 a 的所有变量 从 x 到 a 的所有数字型变量 从 x 到 a 的所有字符型变量

表 1.2 程序的核心语句

语　　句	操作目的
DATA d1;	新建一个数据集 work.d1
INFILE='d:\my1.txt';	外调 D 盘根目录中的 my1.txt 文本数据文件
INPUT v1—v3 (3*2.);	定义三个数字型变量，每个变量的值有两位
CARDS;	宣告读取下面各行的数据

续表

语　　句	操 作 目 的
78 88 89 98 99 87	成百上千的数据行
;	结束数据行。SAS 每条语句结束时也用英文分号";"
DATA d2;	又准备创建一个数据集叫 work. d2
SET d1;	work. d2 数据集是从数据集 work. d1 复制来的
MERGE d1 d2;	横向合并两个数据集,变量可能增加了,但个案数目不变
UPDATE	用一个数据集的数据项更新另一个数据集的数据项
PROC FREQ;TABLE v1;	用具体的统计过程对变量值进行统计或描述
PROC PRINT;	将结果显示在 SAS 的输出窗口

表 1.3　程序中的操作指令

语　　句	作　　用
A=1;	赋值
X=a+b;	累加
IF 表达式;	仅处理满足表达式条件的个案。例如 IF sex=1 THEN c=1;
DELETE v;	删除 v 变量
MISSING a;	遇到字符 a 便视为缺失值
OUTPUT OUT=c1;	输出数据集存入 c1 文件
STOP	中断程序的执行

表 1.4　循环控制语句

语　　句	作　　用
IF sex=2 THEN height * 1.1; ELSE height=height;	对女性的身高赋予 1.1 的权重, 否则就保留原值
GOTO OK; DO DO END; END;	跳转到 OK 语句上执行指令, 外循环开始, 内循环开始, 内循环结束, 外循环结束
LINK RECODE…	临时跳到 LINK 记号处,遇到 RETURN 语句则返回 LINK 处并执行其下一语句
RETURN	返回到 DATA 步执行下面的语句,或返回 LINK 处并执行其下一语句

例 1:GOTO…RETURN 语句组将"RETURN"(返回)到 DATA 步执行它下面的语句。见程序 1.5。

程序 1.5:

```
DATA sj5;
INPUT a b c @@ ;
  IF 4=<a THEN GOTO OK;
```

```
    a=3;
    COUNT+1;
    RETURN;   /* RETURN(返回)到 DATA 步执行它下面的语句 */
  ok:SUMa+a;
CARDS;
3 6 9 10 22 15 12 10 14
;
PROC PRINT;
RUN;
```

运行程序 1.5 产生图 1.6 所示的结果。



图 1.6 GOTO…RETURN 语句的输出

例 2：LINK…RETURN 语句组将返回到 LINK 处并执行其下一语句。有多条 LINK…RETURN 语句，则返回到最后执行的 LINK 语句的下一条上。见程序 1.6。

程序 1.6：将两次成绩中的 D 等成绩升为 C 等。

```
DATA score;
INPUT id test1 $  test2 $  @@ ;
 test=test1;LINK RECODE;
 test=test2;LINK RECODE;
RETURN;
RECODE:IF TEST='d' THEN TEST='c';
RETURN;
CARDS;
01 b c 02 c d 03 a c 04 a d
;
```

```
PROC PRINT;
RUN;
```

运行程序 1.6 产生图 1.7 所示的结果。



图 1.7 用 LINK…RETURN 语句的输出结果

从图 1.7 可以看出,LINK…RETURN 语句将返回到 LINK 处并执行其下一语句,直到将两次成绩中的 D 等成绩升为 C 等为止。

SAS 的管理语句见表 1.5。

表 1.5 SAS 的管理语句

语　　句	作　　用
ARRAY 数组名 [n] [$] [长度] [数组元素]	数组名不仅是变量名,而且是有效的名称,n 为下标,$ 表示字符型变量,例如:ARRAY ab[2] v1 v2;
ATTRIB	指出变量的属性
BY v;	定义分组变量 v
DROP v1;	删除 v1 变量
KEEP v2 v3;	数据集里保留 v2 和 v3 变量
INFORMAT	变量的输入格式
FORMAT sex $;	变量的输出格式
LABEL sex='性别';	变量标签
VALUE sex 1='男' 2='女'	数值标签
LENGTH name $ 15. ;	变量的长度为 15 列字符型
RENAME	改变变量名
RETIAN SUM=0;	初始化变量 SUM 值为 0

管理语句的例子见程序 1.7。

程序 1.7：

```
DATA array;
INPUT a1- a6 b1- b6 ;
  ARRAY test [8] a1- a4 b1- b4 ;
PUT test [4]=   test[5]= ;    /* 日志窗口显示 a4 和 b1 的值 */
CARDS;
1 2 3 4 5 6 101 102 103 104 105 106
11 12 13 14 15 16 111 112 113 114 115 116
;
PROC PRINT;                  /* 显示 a1 至 b6 的值 */
RUN;
```

运行程序 1.7 产生图 1.8 所示的“日志”窗口。

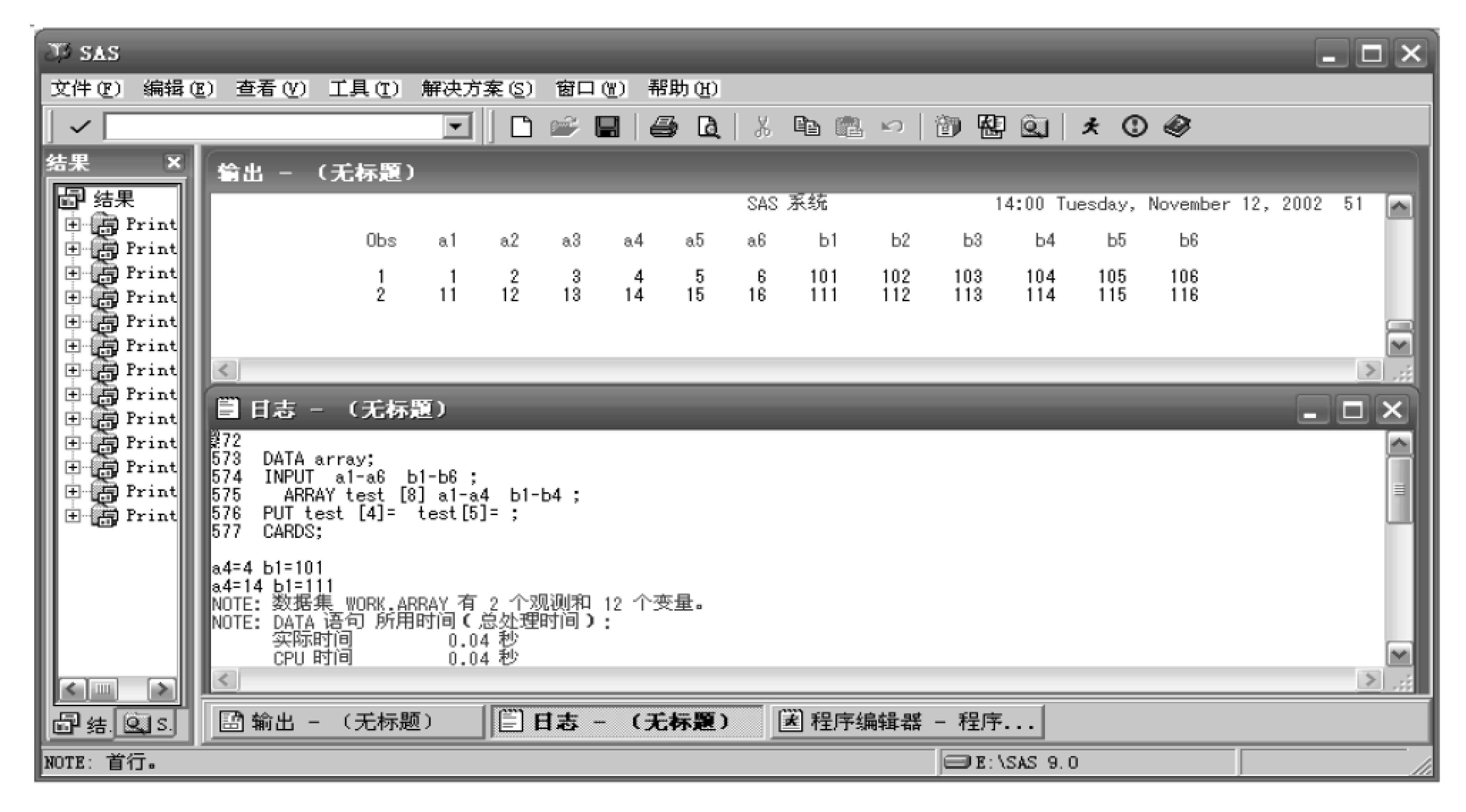


图 1.8　使用数组的技巧

从图 1.8 的“日志”窗口可以看出，确实输出了 a4 和 b1 的值。

习　题　1

1. 什么是观测值(OBS)？
2. 什么是变量(Variable)？
3. 下面的变量名哪些有效？ 哪些无效？

sex、age、v1、location、_ab_、1age、1v、location1、@1、#1、%1、&2

4. 变量有哪些类型？
5. 给程序 a.1 改错。

程序 a.1：

```
DATA sj5;
INPUT a b c @@ ;
  IF 4=<a<15 THEN GOTO OK;
    a=3;
    COUNT+1;
    RETURN; /* RETURN(返回)到 DATA 步执行它下面的语句 */
  OK:SUMa+a;
CARDS;
3 6 9 10 22 15 12 10 14
;
PROC PRINT;
RUN;
```

第 2 章 chapter 2

数据分析的预备知识

SAS 简单编程要用到 DATA 语句、INPUT 语句、LABEL 语句、CARDS 语句、PROC 语句等。

21 DATA 语句

DATA 语句格式如下：

```
DATA _NULL_;    /* 不产生数据集,见图 2.1* /
或 DATA;        /* 产生默认的数据集 WORK.DATA1,见图 2.2* /
或 DATA D1;     /* 产生指定的数据集 WORK.D1,见图 2.3* /
```

SAS 的简单程序见程序 2.1～程序 2.3。

程序 2.1：不产生数据集。

```
DATA _NULL_;
INPUT id a b c@@ ;          /* 用@@表示下面的每行数据包含多个个案 * /
  S=SUM(OF a b c);
  A=MEAN(OF a b c);
PUT S +2 A 10.;             /* 在"日志 (LOG)"窗口显示结果 * /
CARDS;
001 89 91 92 002 91 88 93 003 88 79 95
;
RUN;
```

运行程序 2.1 产生图 2.1 所示的结果。

程序 2.2：产生默认的 WORK.DATA1 数据集。

```
DATA;
INPUT id a b c@@ ;          /* 用@@表示下面的每行数据包含多个个案 * /
  S=SUM(OF a b c);
  A=MEAN(OF a b c);
PUT S A;                    /* 在 LOG 窗口显示输出 * /
CARDS;
```

```
001 89 91 92 002 91 88 93 003 88 79 95
;
RUN;
```

运行程序 2.2 产生图 2.2 所示的结果。

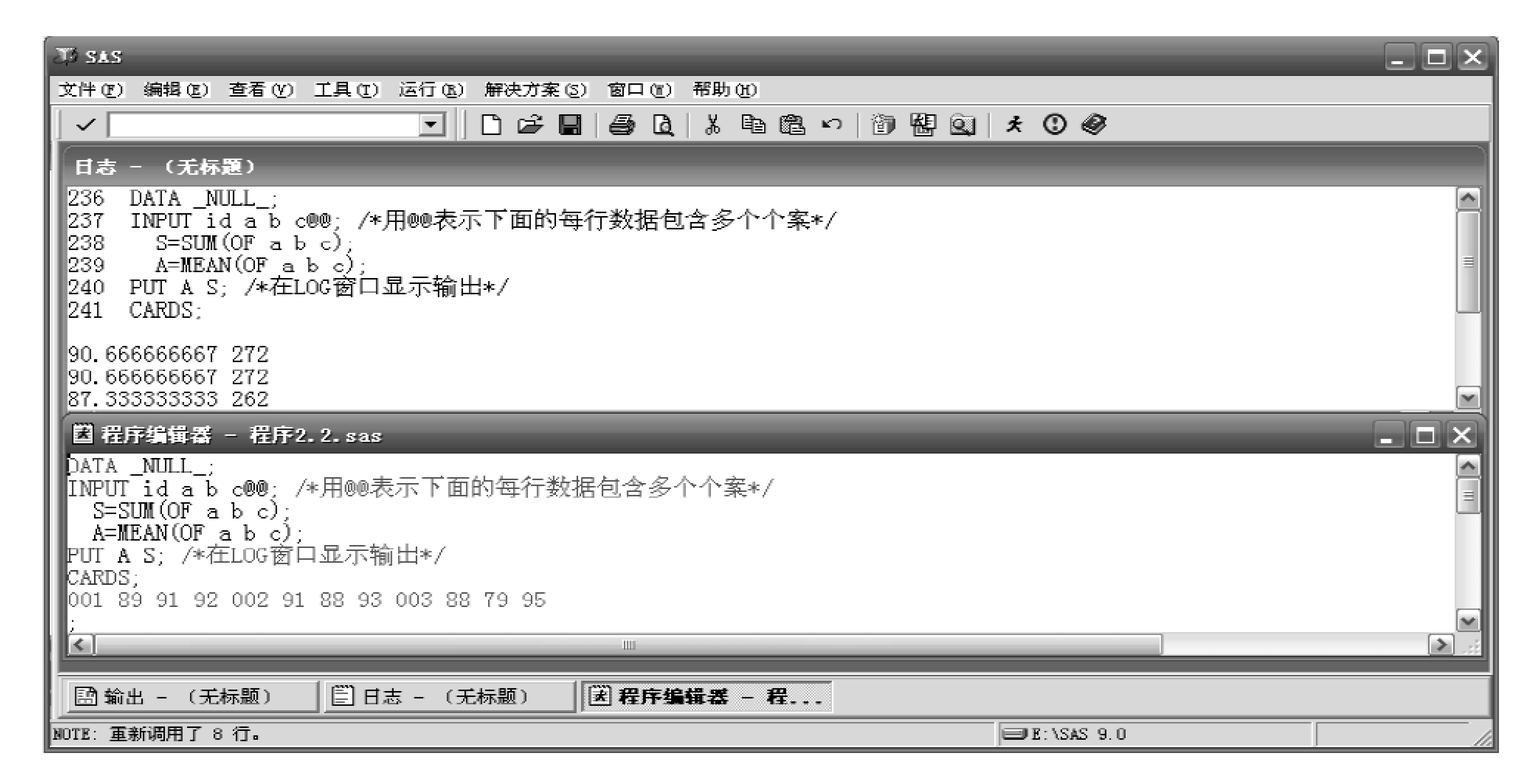


图 2.1 “日志”窗口无数据集名

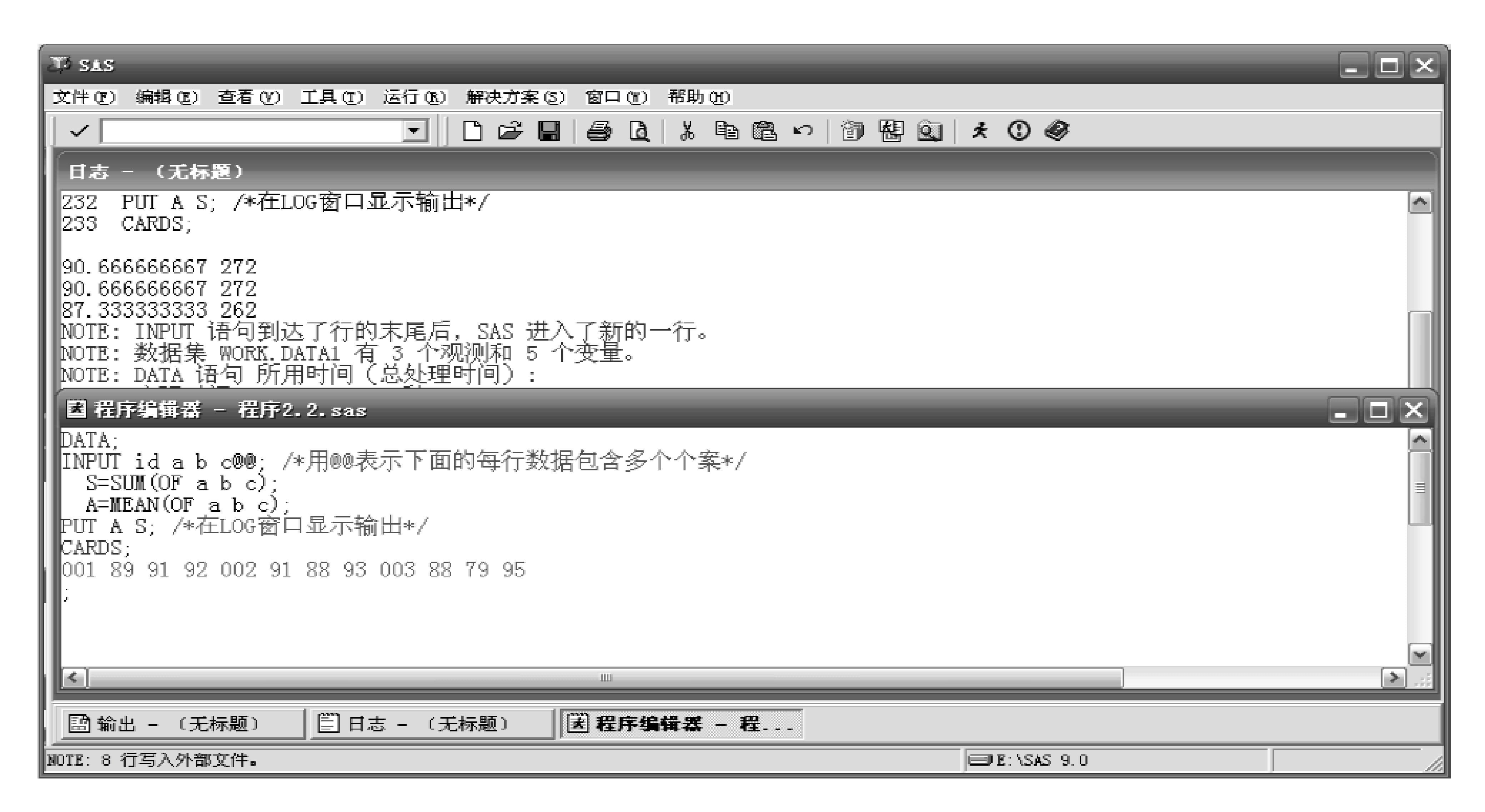


图 2.2 产生数据集 WORK.DATA1

程序 2.3：产生 WORK.D1 数据集。

```
DATA D1;
INPUT id a b c@@ ;          /* 用@@ 表示下面的每行数据包含多个个案 */
  S=SUM(OF a b c);
  A=MEAN(OF a b c);
PUT S A;                    /* 在 LOG窗口显示输出 */
```

```
CARDS;
001 89 91 92 002 91 88 93 003 88 79 95
;
RUN;
```

运行程序 2.3 产生图 2.3 所示的结果。



图 2.3 产生数据集 WORK.D1

22 INPUT 语句

INPUT 语句格式：定义变量名、变量类型、格式、长度等属性。

变量类型分为数字型和字符型两种。变量的格式有自由栏目的数据和固定栏目的数据两种。

1. 3 种自由栏目数据的格式

① INPUT 变量 1 变量 2—变量 n @@；/ * @@表示每行数据含有几个个案 * /。

② INPUT id sex $ age inc1-inc3 @@；/ * 用 $ 表示 sex 为字符型变量 * /。

③ INPUT id sex $ age (inc1-inc3) (3 * 6.1) @@；/ * “3 * 6.1”表示(inc1-inc3)3 个变量各有 6 位数，其中小数位占 1 位 * /。

例 1：格式化输入，见程序 2.4a。

程序 2.4a：

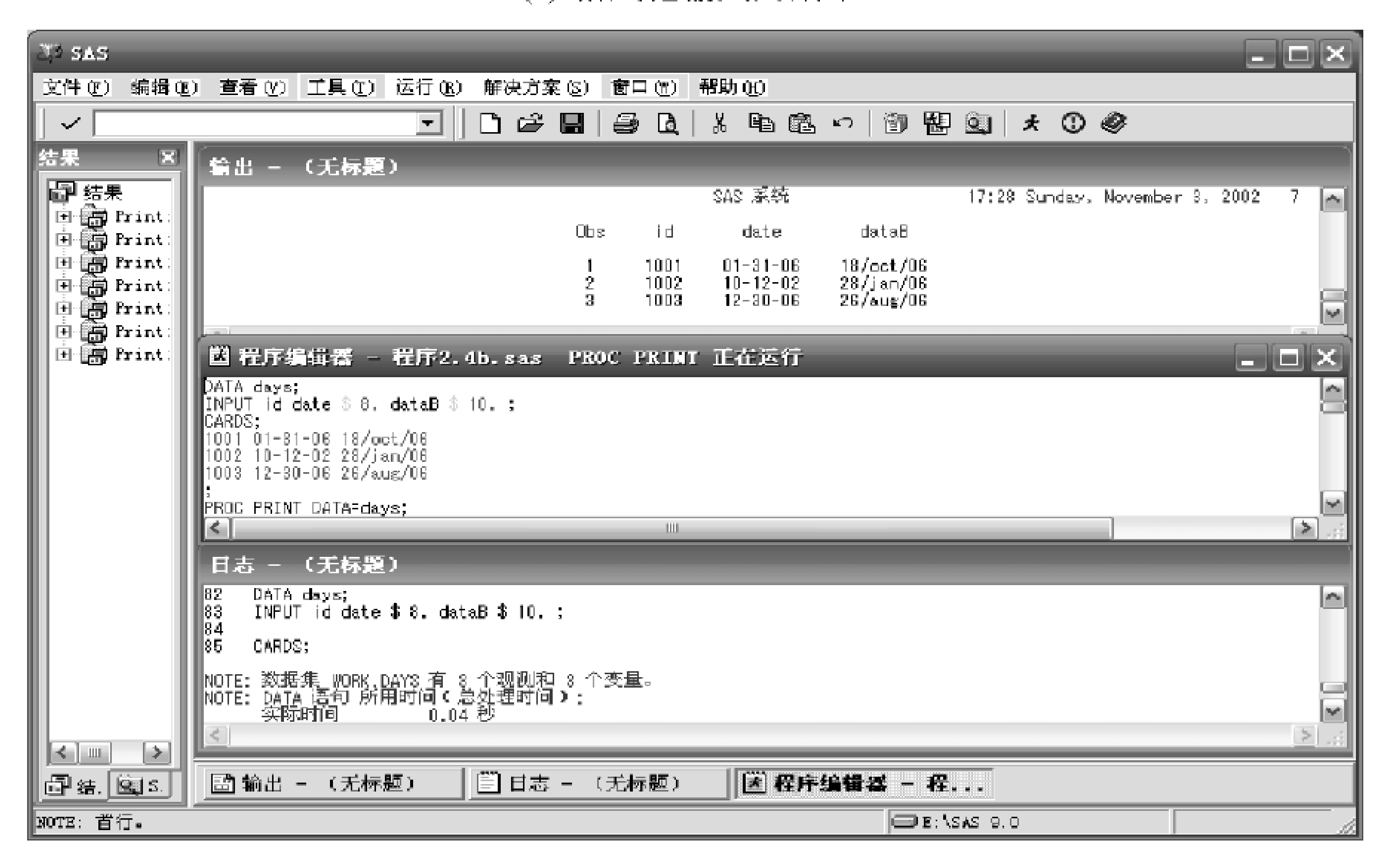
```
DATA;
INPUT id sex $  age 3.2 @@ ;
CARDS;
01 f 150 02 m 162 03 f 170 04 m 162 05 f 168 06 f 161
;
```

```
PROC PRINT;
RUN;
```

运行程序 2.4a 产生图 2.4(a)所示的结果。



(a) 格式化输入的结果



(b) 日期输出

图 2.4　格式化结果及日期的输出

例 2：用格式化输入日期变量，见程序 2.4b。

程序 2.4b：

```
DATA days;
INPUT id date $  8. data $  10. @@ ;
CARDS;
```

```
1001 01-31-06 18/oct/06 1002 10-12-02 28/jan/06
1003 12-30-06 26/aug/06
;
PROC PRINT;
```

运行程序 2.4b 产生图 2.4(b)所示的结果。

注意：

(1) 用格式化输入格式时指针移动是按照人为指定的长度而移动。所以，两个变量的空格也包含在长度内。如果实际的数据长度大于指定的长度“8.”或“n.”长度，则多出的数据被截尾。

(2) SAS 系统初始化数据为“1960 年 1 月 1 日”。因此如果读入的日期是“1960 年 2 月 1 日”时，天数则是 31 天。如果读入的日期是“1960 年 1 月 1 日”以前，则天数是用负值表示。

(3) 指定输出格式时，应该与输入格式相对应。否则，输出的结果难于解释。

例 3：采用 FORMAT 语句对相应时间赋予日期格式。

程序 2.4c：

```
DATA days;
INPUT name $  8. t date8. ;
  FORMAT t YYMMDD8.;
CARDS;
Zhangsan    28oct88
Wangwu      18jan90
Lisi        28nov92
;
PROC PRINT;
RUN;
```

输出结果见图 2.5。



图 2.5 对相应时间赋予日期格式

(4) INPUT V (v1-v10) ($5.); /* 表示(v1-v10)是字符型变量,每个变量长度为5列 */。

(5) INPUT V (v1-v10) (3.); /* 表示(v1-v10)是数字型变量,每个变量长度为3列,没有小数位 */。

INPUT (v1-v10) (3.1); /* 表示(v1-v10)是数字型变量,每个变量长度为3列,小数位占1位 */。

(6) 用行指针#n(或用"/")控读1人多行的数据,见程序2.5及程序2.6。

程序2.5:

```
DATA;
INPUT id1 v1-v3 #2 v4-v6;
CARDS;
1001 70 80 90
75 85 95
1002 60 68 88
76 87 98
;
PROC PRINT;
```

运行程序2.5可产生图2.6所示的结果。

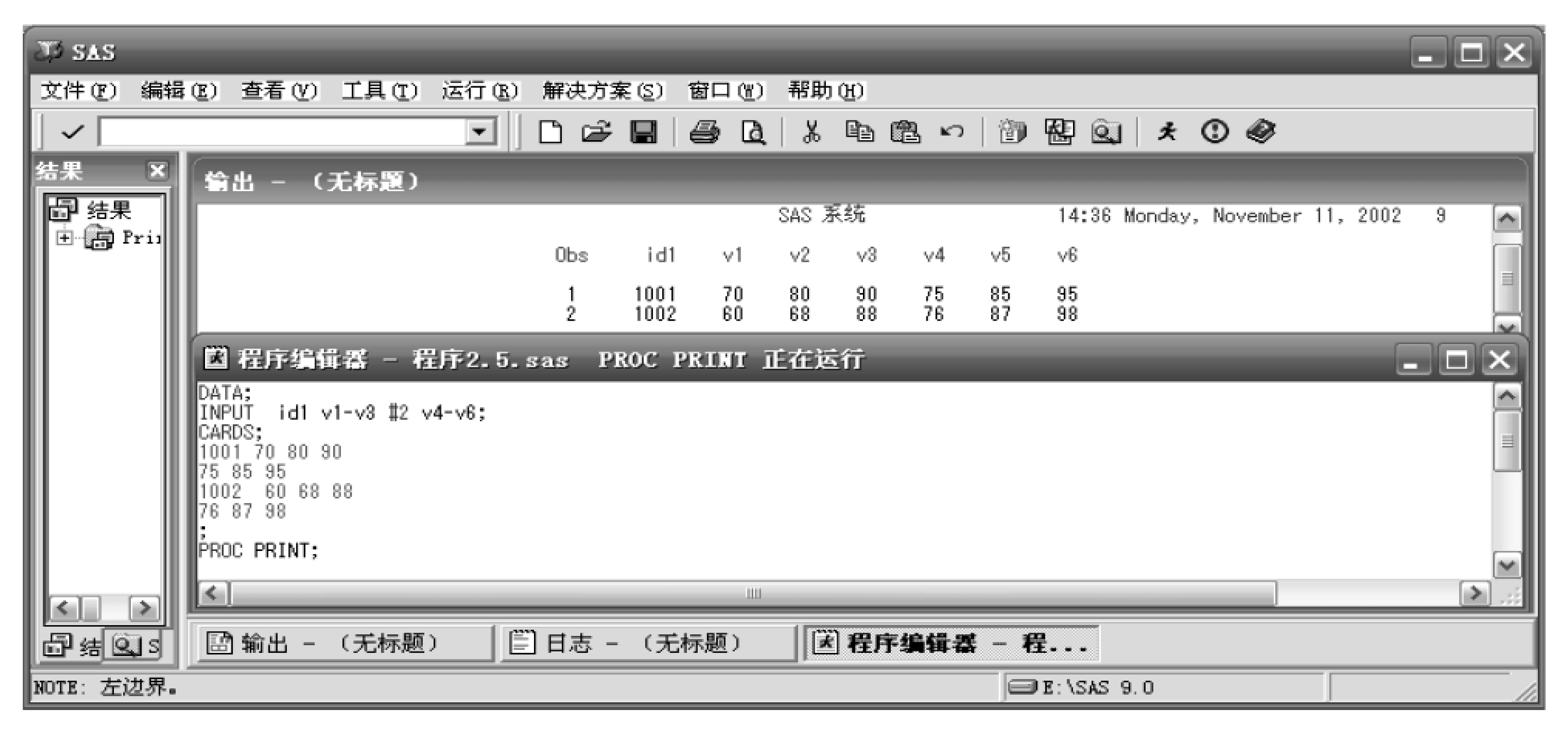


图2.6 用行指针#n控读1人多行数据

程序2.6:用"/"代替#2,也产生图2.6所示的结果。

```
DATA;
INPUT id v1-v3 /v4-v6;
CARDS;
1001 70 80 90 1001 75 85 95
1002 60 68 88 1002 76 87 98
;
PROC PRINT;
```

运行程序 2.6 也能产生图 2.6 所示的结果。

(7) 用列绝对指针@读取自由格式的数据。

• @＃：如跳到 6 列。INPUT id v1-v3 /@6 v4-v6;/ * 见程序 2.7 * /

程序 2.7：

```
DATA;
INPUT id v1- v3/@ 6 v4 v5 v6;
CARDS;
1001 70 80 90
1001 75 85 95
1002 60 68 88
1002 76 87 98
;
PROC PRINT;
```

运行程序 2.7 产生图 2.7 所示的结果。



图 2.7　用列绝对指针@6 读取自由格式的数据

• 用列的相对指针＋n 读取自由格式的数据

＋n：例如向右跳 5 列后读自由格式的数据。

INPUT id v1-v3 ＋5 v4-v6;/ * 见程序 2.8 * /

程序 2.8：

```
DATA;
INPUT id v1- v3 + 5 v4- v6;
CARDS;
1001 70 80 90 1001 75 85 95
1002 60 68 88 1002 76 87 98
;
PROC PRINT;
```

运行程序 2.8 产生图 2.8 所示的结果。



图 2.8　用列的相对指针+n挖掘自由格式的数据

从图2.8的“输出”窗口看,原始数据中每一行的第2个个案号(如1001)自动被略去了。

2. INPUT的固定格式

用INPUT指定变量的固定格式时,每一个变量的值是固定地输入在编辑器的第几列至第几列的栏目位置上,简称固定的栏位。

(1) 一般的固定的栏位例子,见程序2.9。

程序2.9:

```
DATA b;
INPUT id 1-2 sex 4 age 6-7;
CARDS
01 1 28
02 2 38
03 1 45
04 2 36
;
PROC PRINT;
```

程序2.9中,id变量值是固定地输入在编辑器的第1列至第2列的栏目位置上,sex变量值是固定地输入在编辑器的第4列上,age变量值是固定地输入在编辑器的第6列至第7列的栏目位置上。在这里,每个变量的值分别空出1列是为了阅读和辨别。在实际应用中则不必空出1列,以提高数据输入的速度。

(2) 栏位相同的变量可以缩写。

```
INPUT v1 1-3 v2 4-6 v3 7-9;
```

可以简写如下:

```
INPUT (v1-v3) (3* 3.);
```

```
/*“3*”表示有 3个变量。“3.”表示每个变量值的长度都是 3位整数 */
或 INPUT (v1-v3) (3. 3. 3.);
```

(3) 固定格式数据遇到空格则当作缺失值,见程序 2.10。

程序 2.10:

```
DATA;
INPUT id 1-2 sex 4 age 6-7;
CARDS;
01 1 28
02 2 38
03 1
04 2
;
PROC PRINT;
```

运行程序 2.10 产生图 2.9 所示的结果。

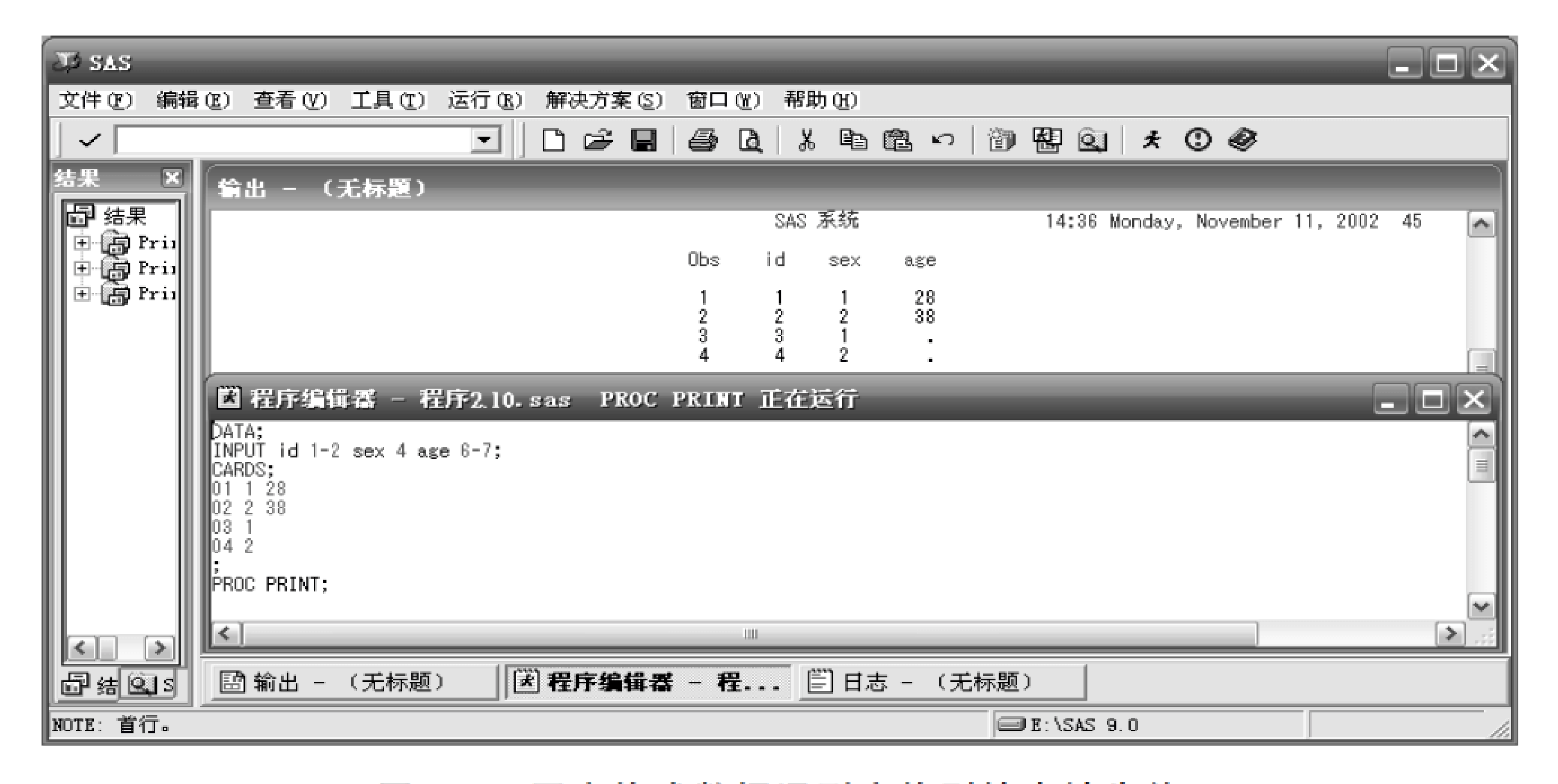


图 2.9 固定格式数据遇到空格则输出缺失值

(4) 固定格式数据的小数位,见程序 2.11。

程序 2.11:

```
DATA;
INPUT id 1-2 sex 4 age 6-8 .1;
CARDS;
01 1 345
02 2 256
03 2 402
04 1 505
05 1.
;
PROC PRINT;
```

运行程序 2.11 产生图 2.10 所示的结果。

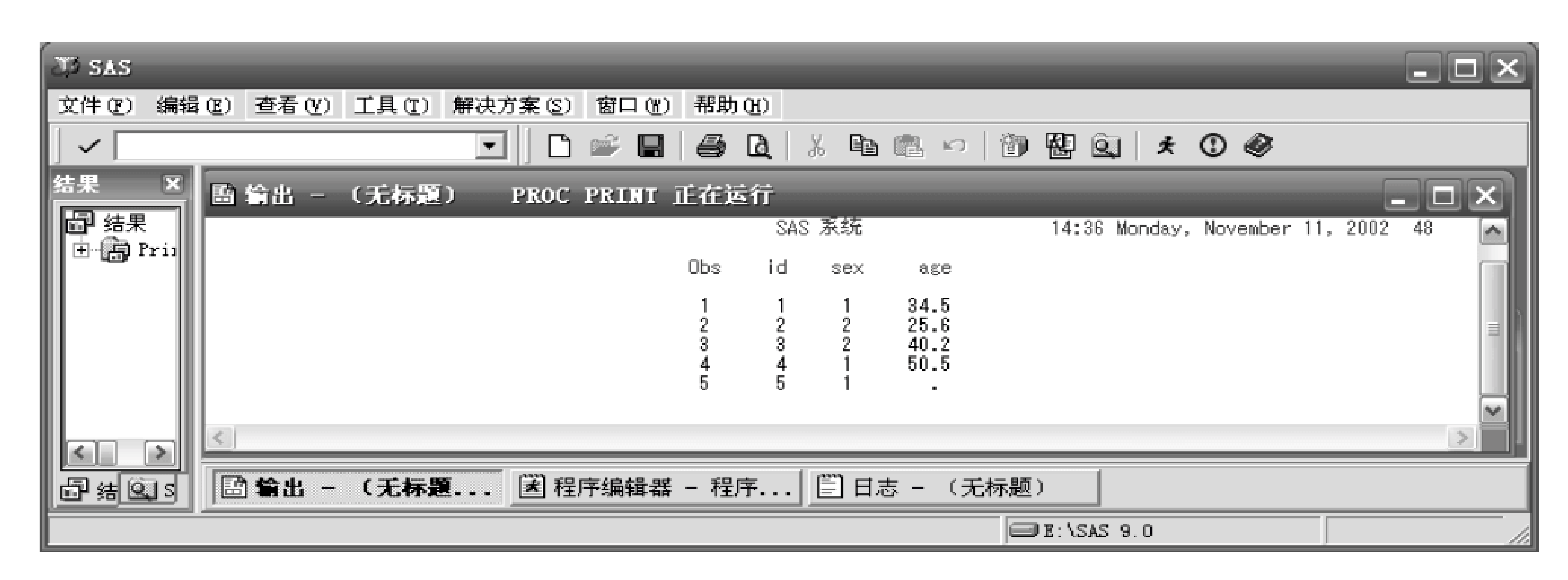


图 2.10　固定格式数据的小数位

从图 2.10 可以看到，年龄已格式化为 34.5 岁等。“.”表示第 5 个人的年龄没有回答而作为缺失值对待。

固定格式小结：

- 固定的栏位必须严格固定地对应，见程序 2.9。
- 固定格式数据遇到空格则当作缺失值，显示“.”，见程序 2.10。
- 每个数字型数据中间不能有空格，否则将被示为是两个变量的值。
- 字符型数据虽然允许空格，但长度在 200 列以内，而且必须用格式化形式。如

INPUT id 1-2 sex 4 name $ 5-19 ；/ * 将 name 变量的值格式化为从 5～19 列字符型长度，共 15 列 * /

或 INPUT id 1-2 sex 4 name $ &；

或 INPUT id 1-2 sex 4 name $ 15.；见程序 2.12。

程序 2.12：

```
DATA b;
INPUT id 1-2 sex 4 age 6-7 name $ 9-19 ;
CARDS;
01 1 28 Zhang san
02 2 38 Li si
03 1 45 Wang wu
04 2 36 Ma liu
;
PROC PRINT;
```

运行程序 2.12 可产生图 2.11 所示的结果。

从图 2.11 可看到，字符型数据虽然允许空格，但必须用格式化形式。

但“INPUT id 1-2 sex 4 **name 5-19 $**；”是错误的，错在 **name 5-19 $** 上。



图 2.11 字符型数据虽然允许空格但必须用格式化形式

23 LENGTH 语句

当字符型数据的长度大于 8 列时，还可用 LENGTH 语句格式化长度。例如在程序 2.12 中的 INPUT 语句前写上一条 LENGTH 语句：

LENGTH name $ 12.;(或 LENGTH name $ 12;)见程序 2.13。

程序 2.13：

```
DATA b;
LENGTH name $  12.;
INPUT id 1-2 sex 4 age 6-7 name &;
CARDS;
01 1 28 Zhang san
02 2 38 Li si
03 1 45 Wang wu
04 2 36 Ma liu
;
PROC PRINT;
```

运行程序 2.13 产生图 2.12 所示的结果。

从图 2.12 的输出窗口，可以验证程序 2.13 与程序 2.12 是等效的。



图 2.12 对程序 2.12 的修改

24 用 LABEL 语句指定变量标签

变量标签使用汉字时最多为 20 个,用字符时最多为 40 个字符,见程序 2.14。

程序 2.14:

```
DATA a1;
INPUT id 1-2 sex 4 age 6-7;
LABEL id='个案号' sex='性别' age='年龄';
CARDS;
01 1 28
02 2 38
03 1
04 2
;
PROC FREQ;
TABLE sex * age;

PROC PRINT DATA=a1;
```

运行程序 2.14 产生图 2.13 所示的结果。

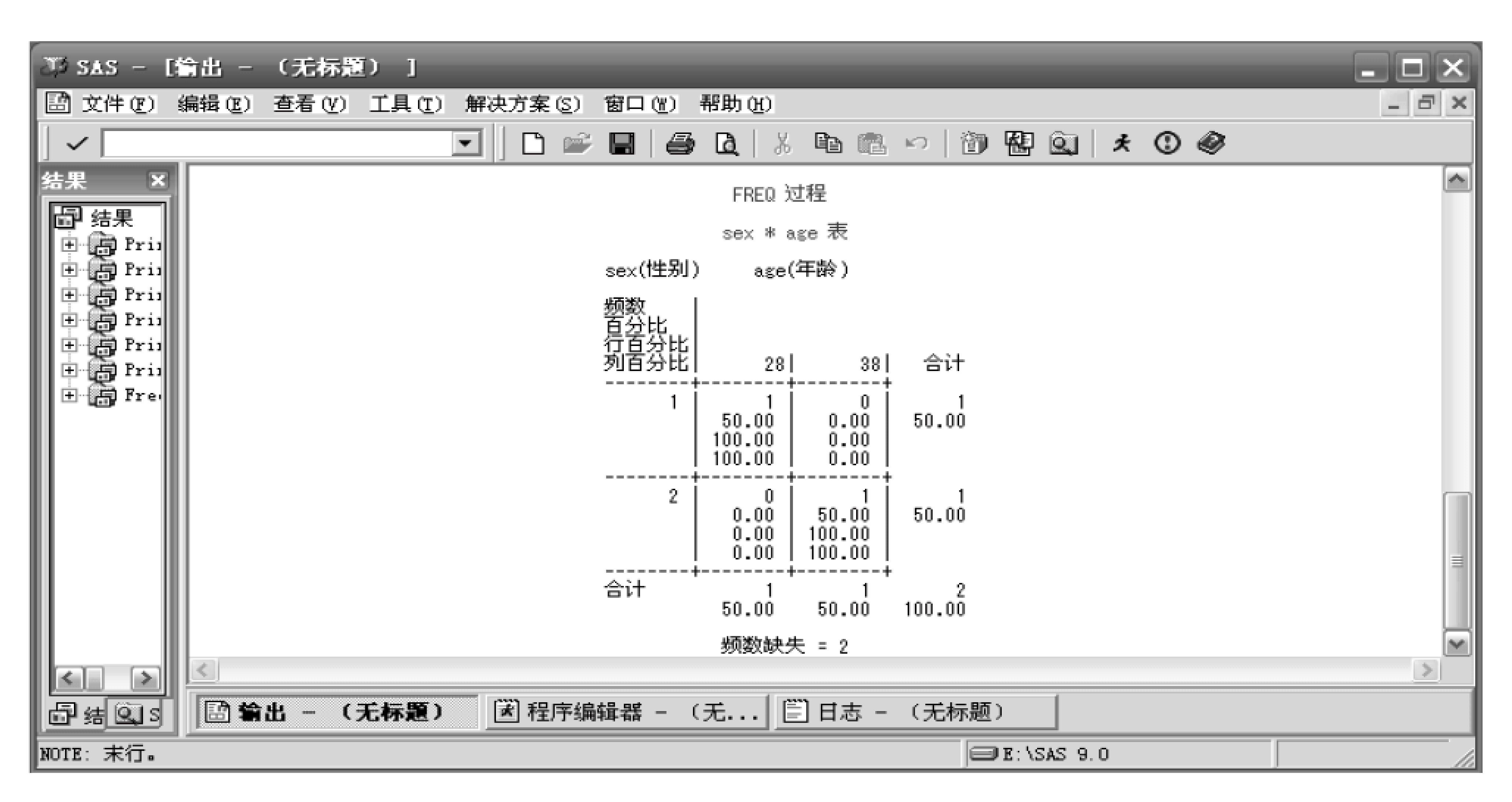


图 2.13 变量成功地汉化了

2.5 用 PROC FORMAT 过程指定数据标签

可用 PROC FORMAT 过程和 VALUE 语句指定变量值(数据)标签。

数据标签使用汉字时最多为 10 个,用字符时最多为 20 个字符,见程序 2.15。

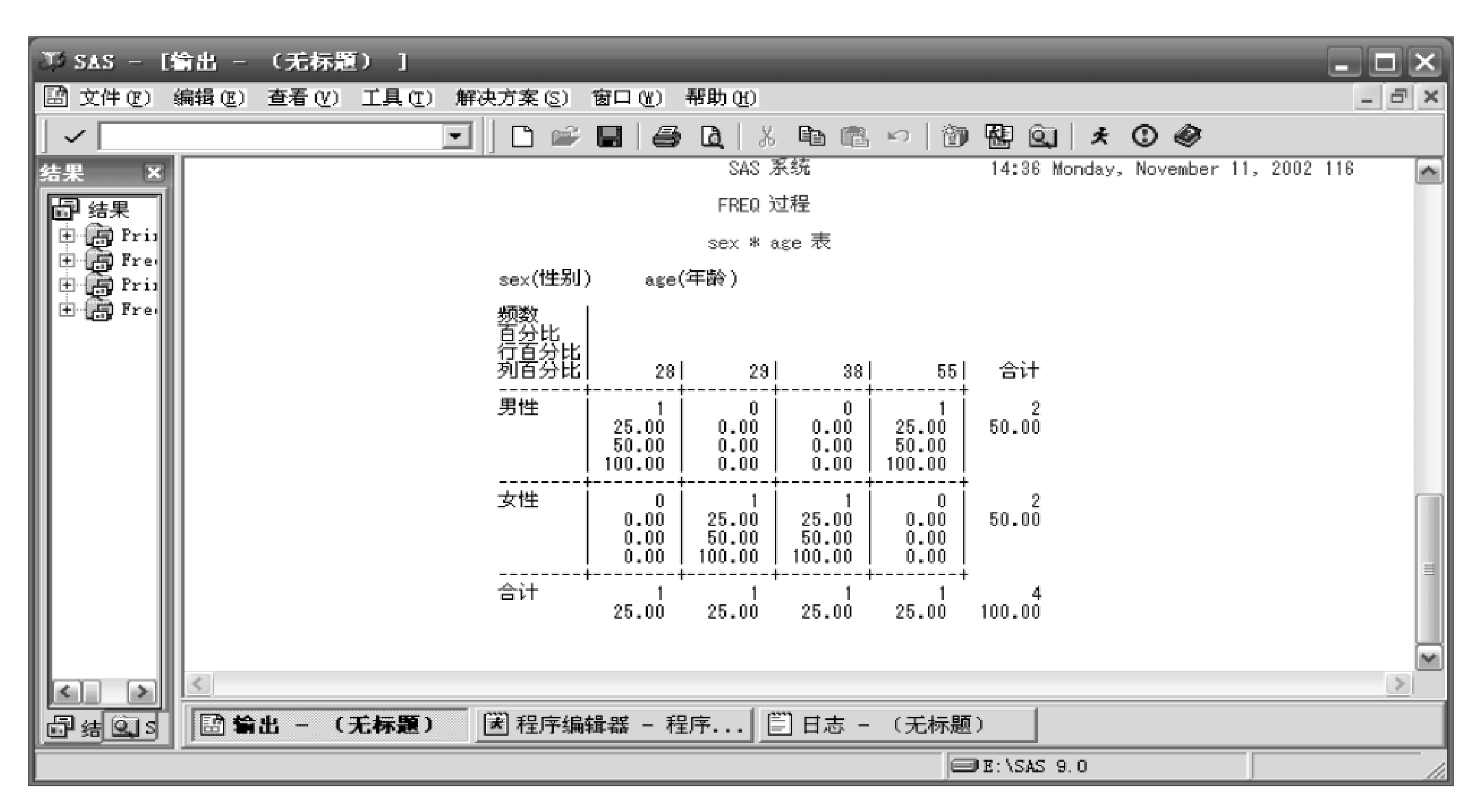
程序 2.15:

```
DATA a1;
INPUT id 1-2 sex 4 age 6-7;
LABEL id='个案号' sex='性别' age='年龄';
CARDS;
01 1 28
02 2 38
03 1 55
04 2 29
;
PROC FORMAT;
  VALUE sexF 1='男性' 2='女性';
 FORMAT sex sexF.;
PROC FREQ;
  FORMAT sex sexF.;
TABLE sex * age;
PROC PRINT DATA=a1;
```

运行程序 2.15 产生如图 2.14 所示的结果。

图 2.14 比图 2.13 更直观,因为又成功地汉化了变量值。

用 FORMAT 语句复制变量值。



sex(性别) 频数/百分比/行百分比/列百分比	age(年龄) 28	29	38	55	合计
男性	1 25.00 50.00 100.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	1 25.00 50.00 100.00	2 50.00
女性	0 0.00 0.00 0.00	1 25.00 50.00 100.00	1 25.00 50.00 100.00	0 0.00 0.00 0.00	2 50.00
合计	1 25.00	1 25.00	1 25.00	1 25.00	4 100.00

图 2.14　变量值又成功地汉化了

例如：FORMAT sex sexF.；(详见程序 2.15)

26　用 MISSING 语句宣告缺失值

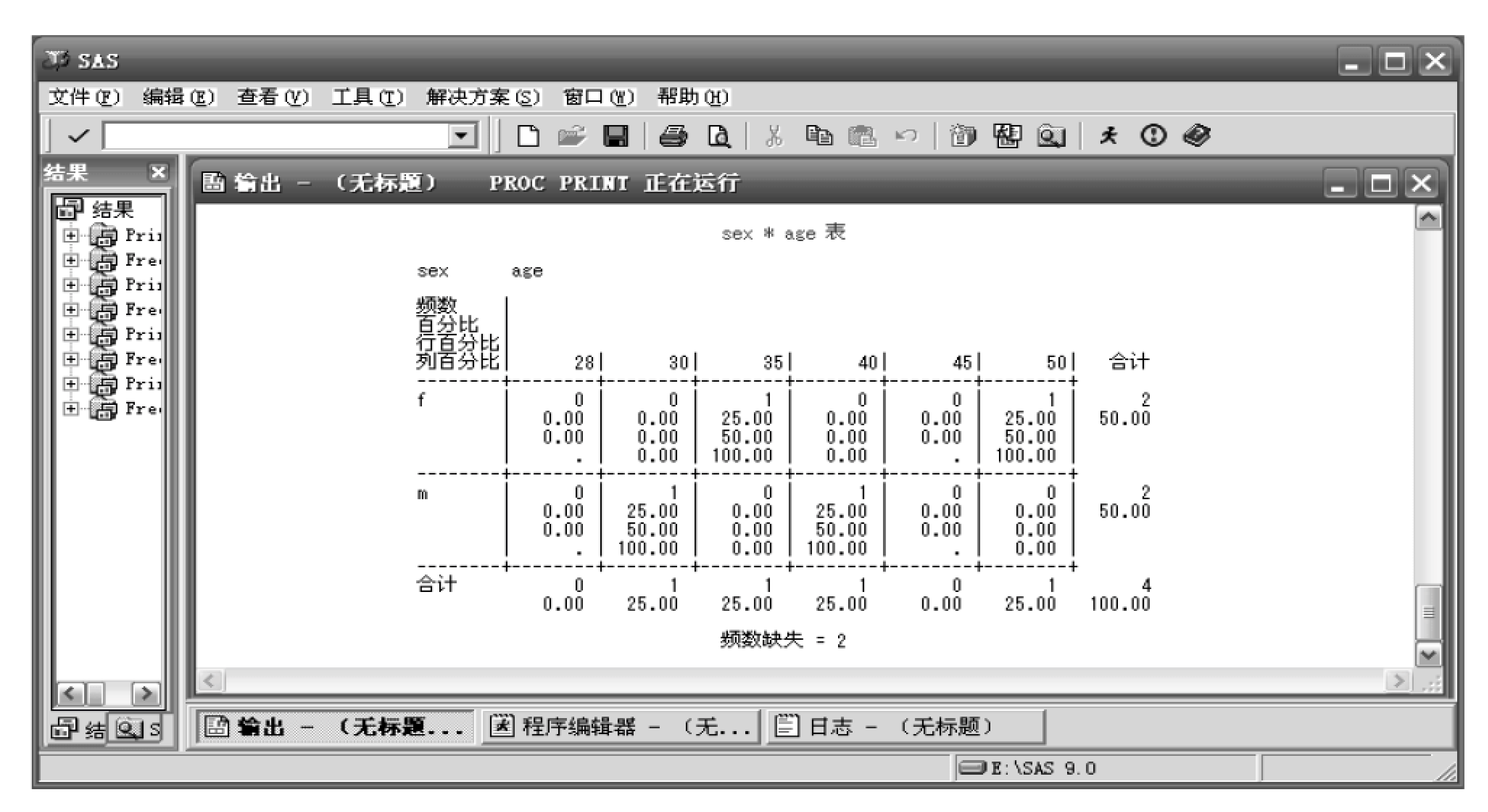
被访者不愿意回答或漏答的数据可以作为缺失值处理。数字型数据的缺失值是用“.”输入和表示的。字符型数据的缺失值可以用“A”表示访问时家里没有人,用“R”表示访问时拒答,详见程序 2.16。

程序 2.16：

```
DATA m1;
MISSING A R;/* 由于此语句在微型机上行不通,在下面的数据行中仍用'.'"表示缺失值"*/
INPUT id sex $  age @@ ;
/* 在下面的数据行中只能用'.'表示缺失值;*/
CARDS;
01 m 40 02 f 50 03 m 30 04 . 28 05 f 35 06 . 45
;
PROC FREQ;
TABLE sex*age;
PROC PRINT DATA=m1;
```

运行程序 2.16 产生图 2.15 所示的结果。

说明：“MISSING A R;”语句在微型机上行不通,所以建议数据还是采用数字型的好。



sex * age 表

sex / age 频数 百分比 行百分比 列百分比	28	30	35	40	45	50	合计
f	0 0.00 0.00 .	0 0.00 0.00 0.00	1 25.00 50.00 100.00	0 0.00 0.00 0.00	0 0.00 0.00 .	1 25.00 50.00 100.00	2 50.00
m	0 0.00 0.00 .	1 25.00 50.00 100.00	0 0.00 0.00 0.00	1 25.00 50.00 100.00	0 0.00 0.00 .	0 0.00 0.00 0.00	2 50.00
合计	0 0.00	1 25.00	1 25.00	1 25.00	0 0.00	1 25.00	4 100.00

频数缺失 = 2

图 2.15 有缺失值的显示结果

27 注释语句

注释语句有以下几种格式：

* ……；(一般用在语句行的上一行)

或/ * 注释内容 * / (一般用在语句行的行尾)

见程序 2.16 中语句“CARDS;”前面的“/ * 在下面的数据行中只能用‘.’表示缺失值 * /”。

28 创建新变量

语句格式：

Y=x1 运算符号 变量名;(例如 y=x1+x2;)

Y=x1 运算符号 数值;(例如 yinc=mincl * 12;)

程序 2.17：从月收入中创建年收入的变量。

程序 2.17：

```
DATA a2;
INPUT id 1-2 sex 4 age 6-7 mincl 9-13 .1;
LABEL id='个案号' sex='性别' age='年龄' mincl='月收入';
yinc=mincl * 12;
CARDS;
01 1 28 45000
02 2 38 36000
```

```
03 1 55 30000
04 2 29 25000
;
PROC FORMAT;
  VALUE sexF 1='男性' 2='女性';
 FORMAT sex sexF.;
PROC FREQ;
TABLE sex * age;
PROC PRINT;
```

运行程序 2.17 产生图 2.16 所示的结果。



图 2.16 由程序 2.17 所产生的运行结果

从图 2.16 可看到，已经新创建了一个 yinc 变量。

29 缺失值不参与运算

程序 2.18 标明缺失值不参与运算。

程序 2.18：

```
DATA a3;
INPUT id score1 score2 @@ ;
Score1= .;                /* 结果为 . */
Score2= score1+ 5;
Score= SUM(score1,5);     /* 结果为 5 */
CARDS;
01 80 90 02 78 88 03 68 79
```

```
;
PROC PRINT;
```

运行程序 2.18 产生图 2.17 所示的结果。

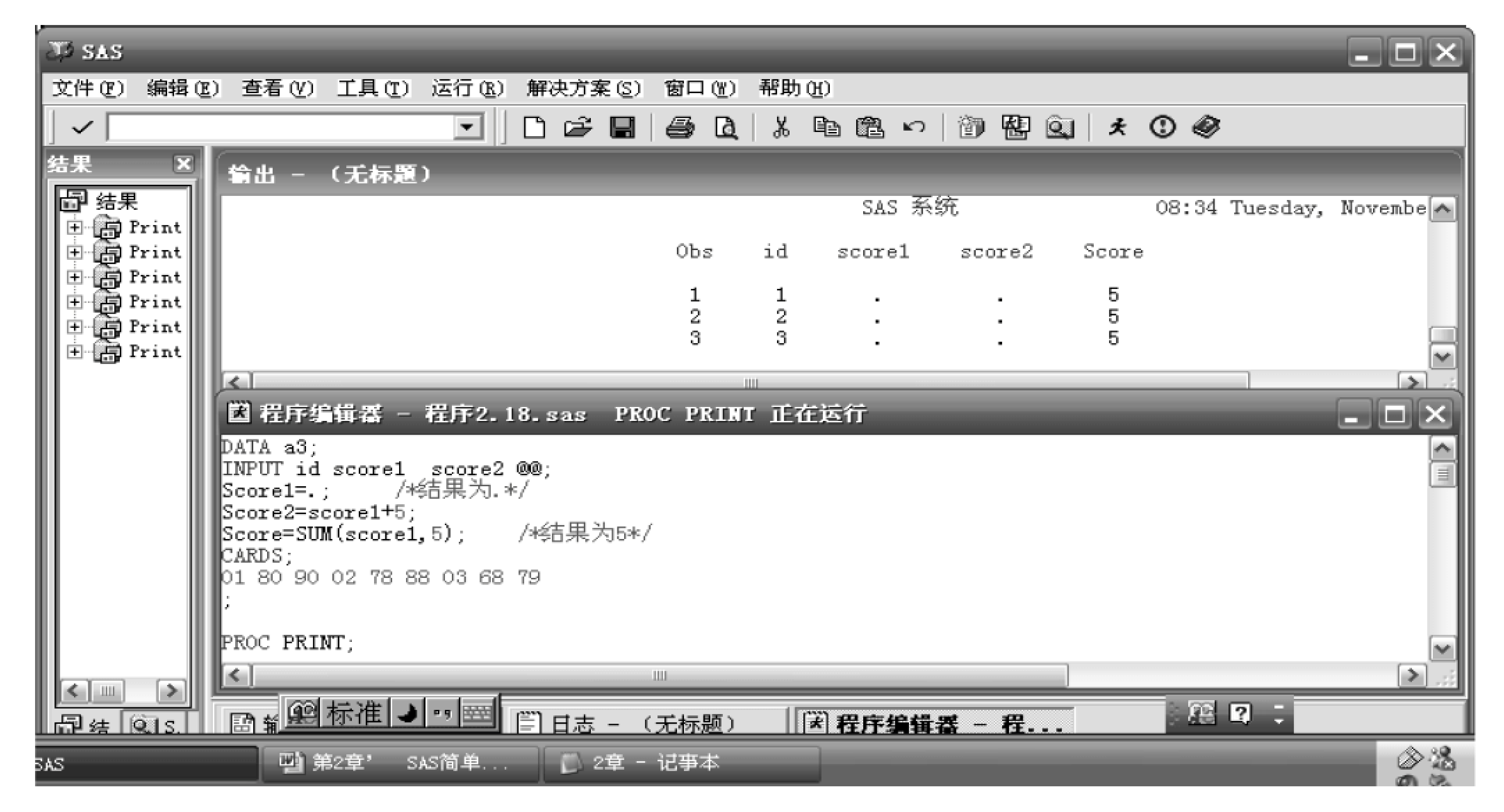


图 2.17 缺失值不参与运算

从图 2.17 可以看到，有缺失值则不累加。

2.10 求和语句

语句格式：

```
S=SUN(a,b);
Y+x;
```

程序 2.19：

```
DATA;
INPUT x1 x2 @@ ;
    Y+x1; /* 累加 * /
CARDS;
20 30 40 20 25 18
;
PROC PRINT;
```

运行程序 2.19 产生图 2.18 的上半图。

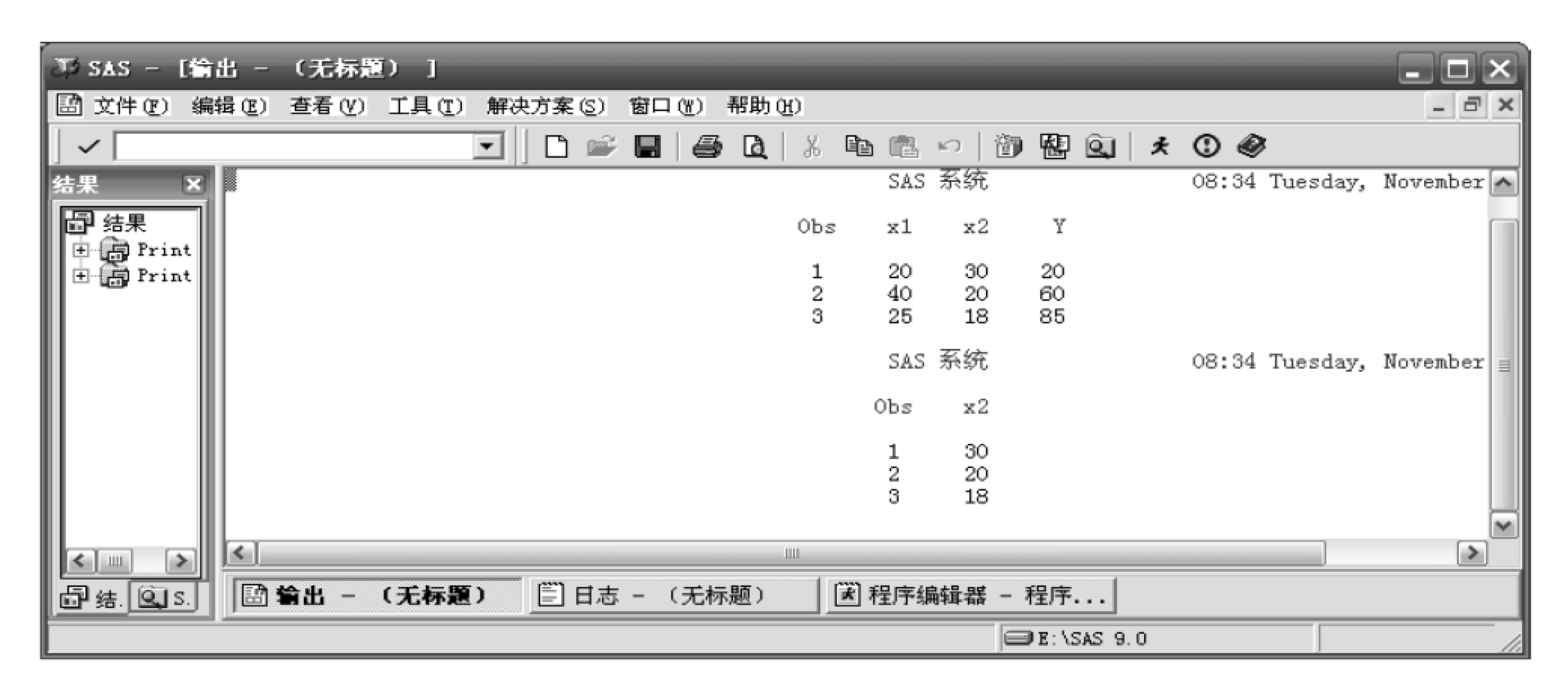


图 2.18 用"Y＋x1;"语句累加

2.11 删除变量

语句格式:

DORP v; (例子见程序 2.20)

程序 2.20:

```
DATA x;
INPUT x1 x2 @@ ;
Y+ x1; /* 累加,产生图 2.18的上半图 * /
CARDS;
20 30 40 20 25 18
;
PROC PRINT;
DATA d;
 SET x;
 t= SUM(OF x1 x2);
 DROP x1; /* 删除变量 x1,产生图 2.18的下半图 * /
KEEP x2;
PROC PRINT;
RUN;
```

运行程序 2.20 产生图 2.18 所示的结果。

2.12 用 INFILE 语句读取外部文件的数据

当样本量有成百上千份问卷时,应该单独输入数据并存储为数据文件。用时再用 INFILE 语句读取此外部文件的数据。

语句格式：

INFILE 'D:\my1.txt'; 例见程序 2.21。

先用附件中的“记事本”作为编辑器，在 D 盘的根目录上建立'my1. txt'，即'D:\my1. txt'。下面是‘my1. txt’的部分内容：

```
0001 2 26 3000 0002 2 30 3200 03 1 31 3300 04 2 29 2990
…
```

然后用程序 2. 21 的 INFILE 语句调用'D:\my1. txt'中的数据进行统计。

程序 2. 21：

```
DATA s;
INFILE 'D:\my1.txt';
INPUT id sex age mincl @@ ;
  Y=mincl * 12;
PROC PRINT;
RUN;
```

运行程序 2. 21 产生图 2. 19 所示的结果。



图 2. 19 用“INFILE 'D:\my1. txt';”语句成功地调用外部数据

习 题 2

1. 指出下列命令的作用。

(1) DATA _NULL_;

(2) DATA;

(3) DATA D1;

2. 自由栏目数据有哪些格式？

3．试举一个格式化输入的例子。
4．试举一个格式化输入日期变量的例子。
5．试举一个格式化输出日期格式的例子。
6．固定栏目数据有哪些格式？
7．试用“/”控读一人两行的数据。
8．试举一个固定栏位的程序。
9．试举一个用 LABEL 语句指定变量标签的例子。
10．试举一个用 VALUE 语句指定数值标签的例子。
11．试举一个创建新变量的例子。

第 3 章 chapter 3

数据的跳转与循环

要从整体数据中挖掘部分目标数据，就要根据某些条件是否成立。例如，当条件成立时转移到另外语句上临时执行其他一些指令，执行完毕便返回到刚才的位置上。这种转移形式则称为“跳转”。

用于跳转的语句有如下所示的几种：

- IF <表达式> THEN 语句；/ * 如果表达式的条件成立，则执行它后面的语句 * /
- IF <表达式> THEN 语句 1；ELSE 语句 2；
  / * 如果表达式的条件成立，则执行它后面的语句 1；否则就执行语句 2 * /

3.1 IF 语句

3.1.1 IF THEN 语句

1. 语句格式

```
IF <表达式> THEN 语句；
```

此格式表示，如果条件成立时则反馈“1”，并继续执行“THEN 语句”的语句指令。

如果条件不成立则反馈“0”，不执行“THEN 语句”的语句指令，而跳转到下面一个语句上加以执行。

2. 例子

例 1：计算 A、B、C 这 3 次考试成绩之和，如果和大于 270 分，则显示出学号及总成绩，见程序 3.1。

程序 3.1：

```
DATA score1;
INPUT id a b c;
TOTAL= SUM(OF a b c);
IF TOTAL> 270 THEN PUT id total;
```

```
CARDS;
001 89 91 92 002 91 88 93 003 88 79 95
;
RUN;
```

运行程序 3.1 后在“日志(LOG)”窗口显示：1 272(第 1 位学生总分 272 分，其他没有达到 270 分)，见图 3.1。

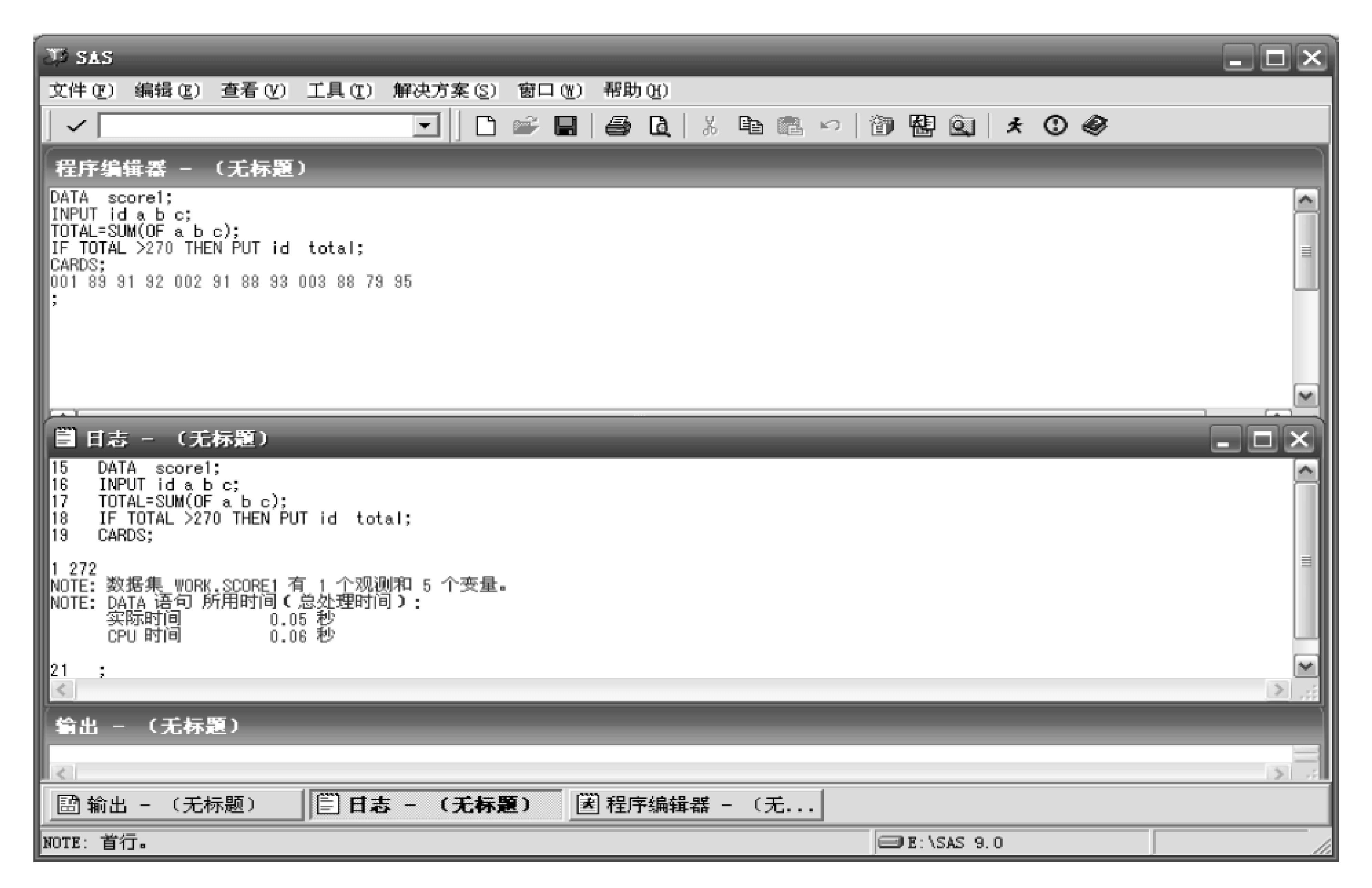


图 3.1 学生 3 门功课的总分

例 2：对于 A、B、C 这 3 次考试成绩每门大于 90 分的，则显示出学号及总成绩，见程序 3.2。

程序 3.2：

```
DATA score2;
INPUT id a b c @@ ;
IF a> 90 OR b> 90 OR c> 90 THEN PUT id a b c;
CARDS;
001 89 91 92 002 91 88 93 003 88 79 95
;
RUN;
```

运行程序 3.2 后在“日志(LOG)”窗口显示出 3 行的结果，见图 3.2。

例 3：对于 A、B、C 这 3 次考试成绩平均大于 85 分的，则显示出学号及总成绩，见程序 3.3。

程序 3.3：

```
DATA score3;
INPUT id a b c@@ ;
```

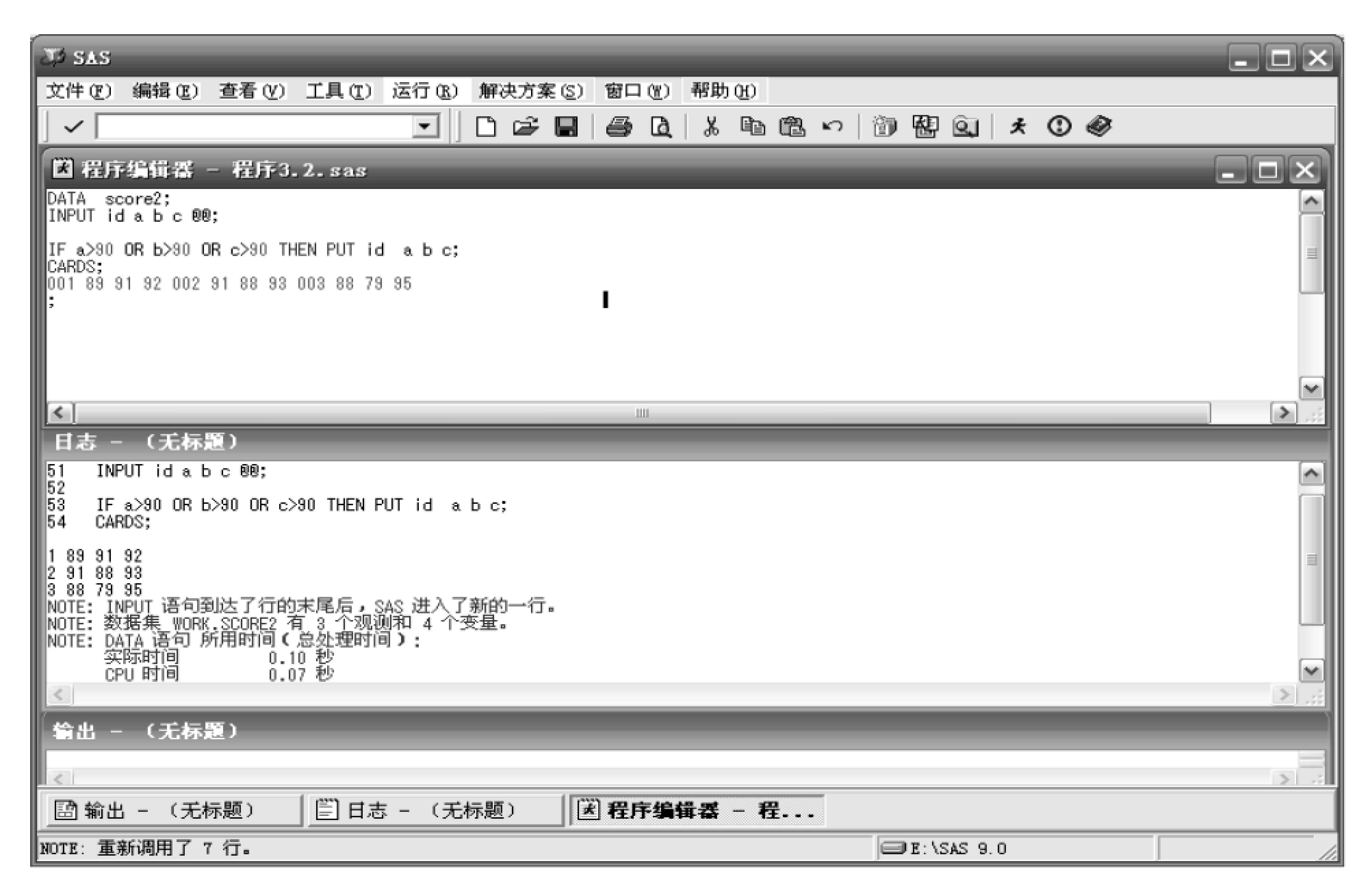


图 3.2　学生 3 门功课大于 90 分的则显示其结果

```
 M=MEAN(OF a b c);
IF M>85 THEN PUT id M;
CARDS;
001 89 91 92 002 91 88 93 003 88 79 95
;
RUN;
```

运行程序 3.3 后在"日志(LOG)"窗口显示出 3 行的结果，见图 3.3。

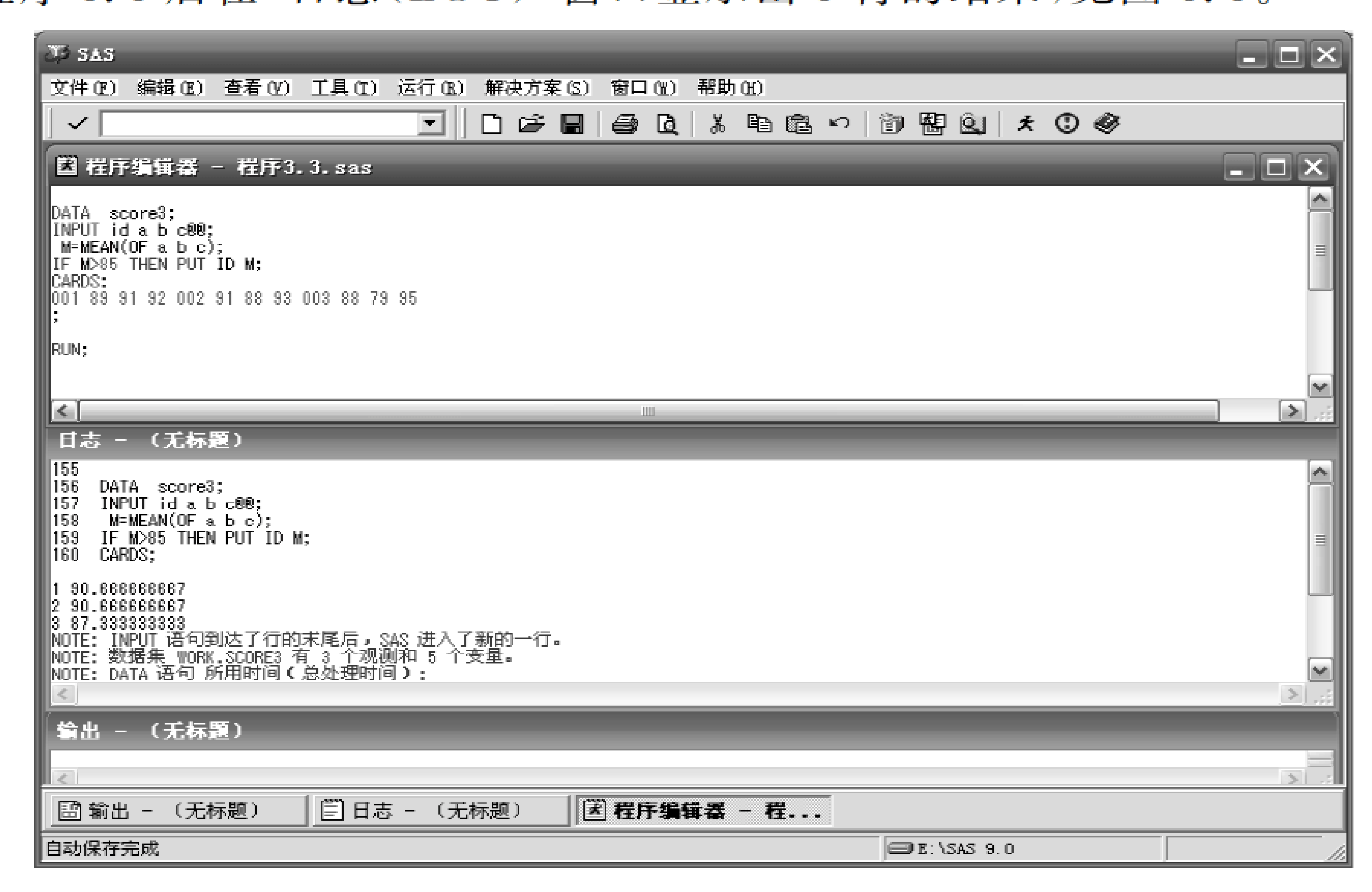


图 3.3　A、B、C 3 次考试成绩平均大于 85 分的个案

3.1.2 IF THEN/ELSE 语句

1. 语句格式

```
IF <表达式> THEN 语句 1;
ELSE 语句 2;
```

2. 说明

此语句表示：如果表达式的条件成立，则执行它后面的语句 1；否则就执行语句 2。

3. 例子

例 4：在学历调查中有以下 4 种情况。
1—高中以下，2—大专，3—本科，4—硕士以上。
若统计大专以下的人数，则建立两个变量：da、db。那么，语句将是：

```
IF EDU LE 2 THEN dz=1; ELSE dz=0;
IF EDU GT 3 THEN db=1; ELSE db=0;
```

命令语句见程序 3.4。
程序 3.4：

```
DATA edu1;
INPUT id edu @@ ;
IF edu LE 2 THEN dz=1; ELSE dz=0;
IF edu GE 3 THEN db=1; ELSE db=0;
CARDS;
001 2 002 3 003 4 004 2 005 3 006 4
;
PROC PRINT DATA=edu1;
RUN;
```

运行程序 3.4 后在输出(OUTPUT)窗口显示出结果，见图 3.4。

IF…THEN 后面只能用一个语句。如果条件满足要执行多个语句，则应采用 DO…END 语句。例如，如果 a=3，则将 3 改为 4，并且显示出该个案号。这时可采用以下几条语句：

```
IF a=3 THEN DO;
a=4;
PUT a;
END;
```

例 5：IF…THEN/ELSE 可以嵌套，见程序 3.5。
程序 3.5：

```
DATA I1;
```

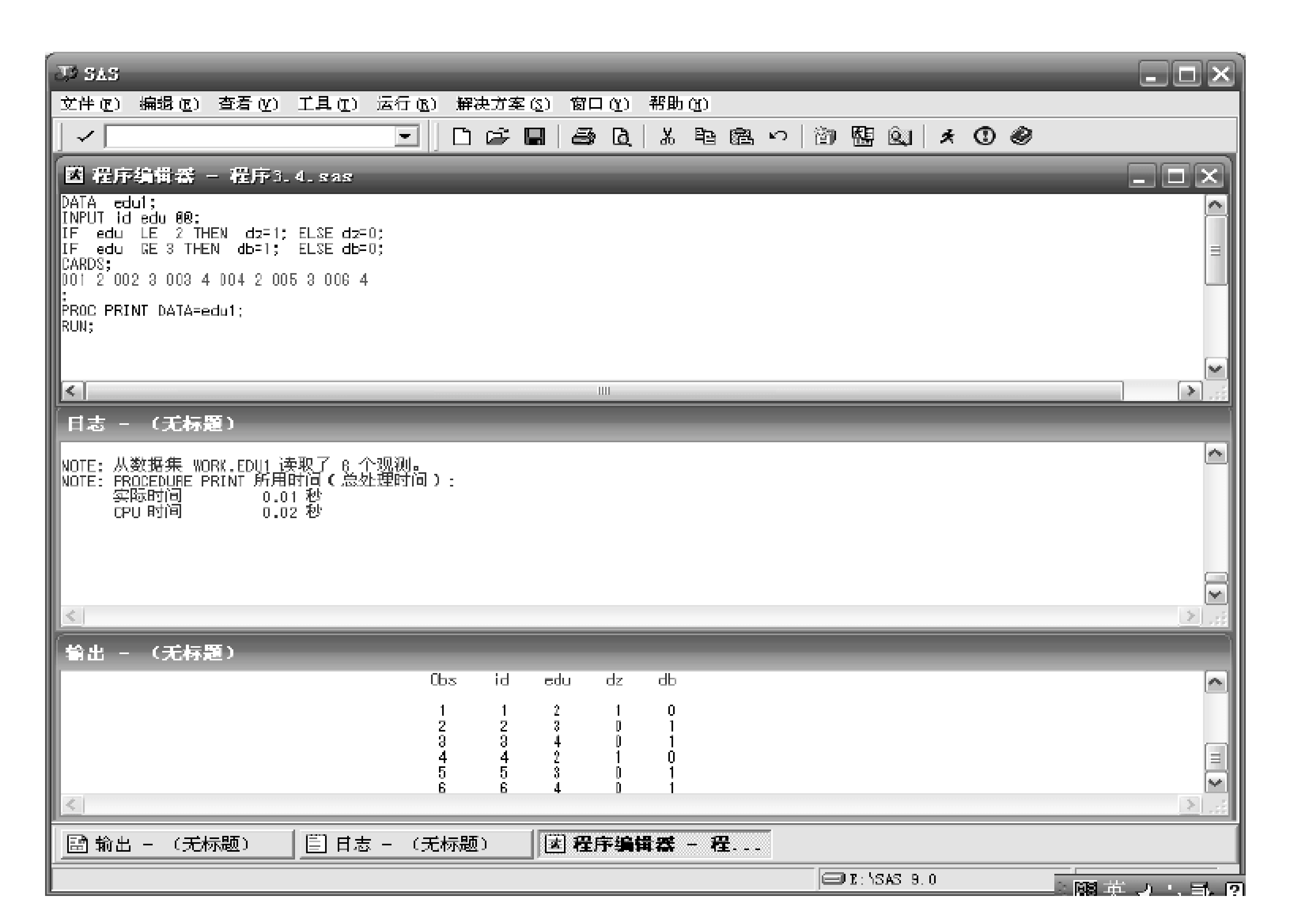


图 3.4 学历调查

```
INPUT id1 id2 a b @@ ;
IF a=1 THEN DO;
    IF b=1 THEN PUT 'a=1 & b=1的个案' ELSE PUT id 'a=1 & b=0的个案';
    END;
 ELSE PUT id 'a不等于 1的个案';
CARDS;
001 01 1 1 002 01 1 0 003 02 0 1 004 02 0 0
;
PROC PRINT;
```

运行程序 3.5 在输出(OUTPUT)窗口显示出结果,见图 3.5。



图 3.5 IF…THEN/ELSE 的嵌套结果

3.2 GO TO语句

GO TO 语句(或 GOTO 语句)让 SAS 立即跳向 GO TO 所指的语句,并从该语句开始执行。GO TO 语句与跳向的目标必须在同一个 DATA 步中。目标可用 OK 或 LABEL 等其他标号。

1. 命令格式

```
GOTO OK;
…
OK:x+1;
```

2. 例子

例 6:GO TO 语句用在 IF…THEN 的 THEN 后面表示跳转,见程序 3.6。

程序 3.6:

```
DATA G1;
INPUT id a b @@ ;
IF a>=5 AND a<=8 THEN GOTO OK;
    a=6;
  COUNT+1;
OK:SUMa+b;
CARDS;
001 8.8 5.5 002 7.5 8.5 003 6.5 7.5
;
PROC PRINT;
```

运行程序 3.6 后在输出(OUTPUT)窗口显示出结果,见图 3.6。

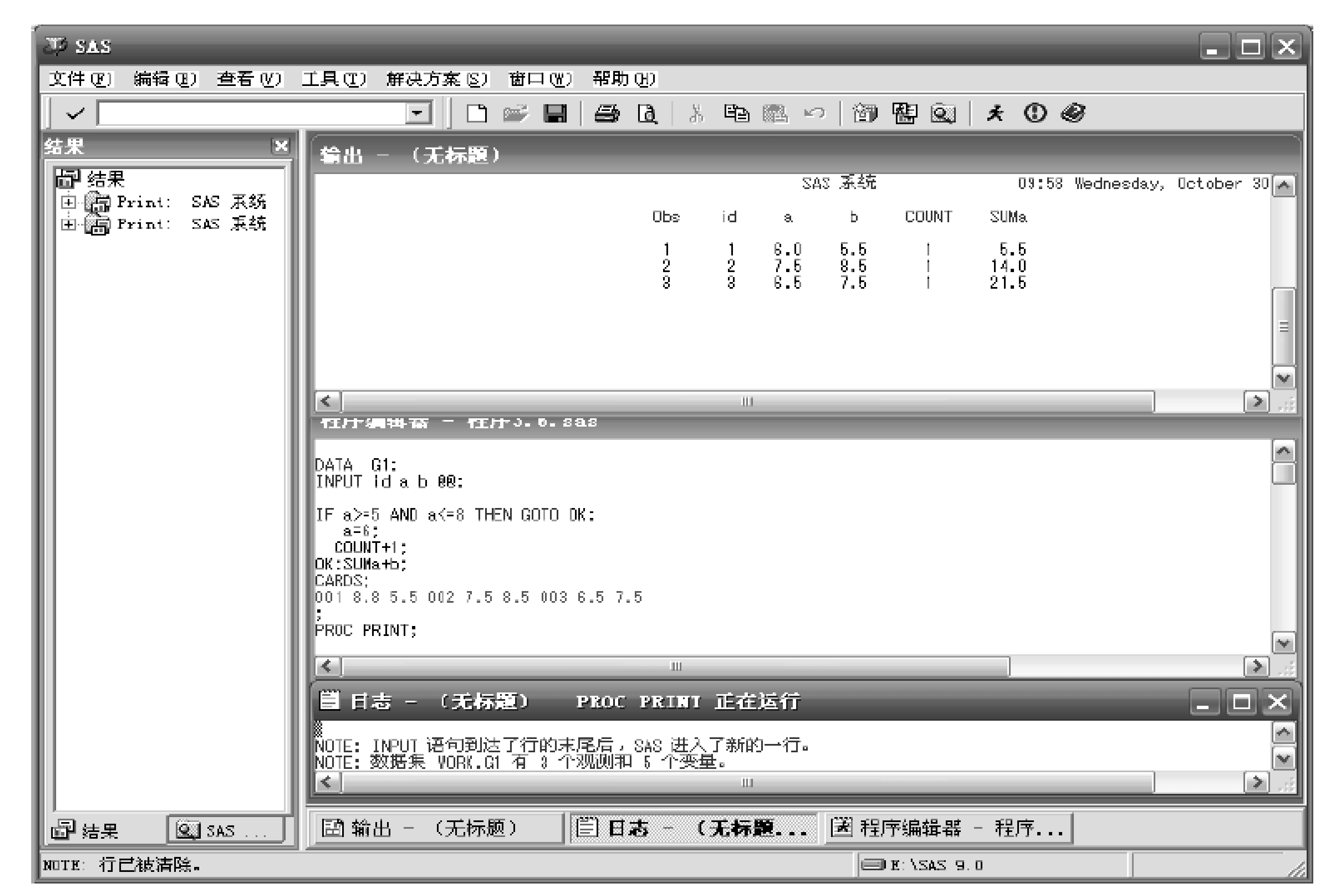


图 3.6 GO TO 语句的输出

3.3 LINK语句

用 LINK 语句让 SAS 指向临时要执行的子程序语句。当遇到 RETURN 语句时，则返回到刚才 LINK 语句的后面语句上继续向下执行。

1. 命令格式

```
LINK 目标; /* 目标是任意的一个 SAS 命令。如 COMPUTE、RECODE 等 * /
 LINK: /* 可以嵌套 10 层 LINK…RETURN 语句 * /
  RETURN
RETURN;
```

2. 例子

例 7：将 3 次考试成绩中的 3 分改为 4 分，见程序 3.7。

程序 3.7：

```
DATA a1;
INPUT id t1 t2 t3 @@ ;
     T=t1; LINK COMPUTE;
     t1=T;
     T=t2; LINK COMPUTE;
     t2=T;
     T=t3; LINK COMPUTE;
     t3=T;
  COMPUTE:IF T=3 THEN T=4;
RETURN;
CARDS;
001 3 4 5 002 4 3 5 003 3 4 4
;
PROC PRINT;
```

说明：为了每次给成绩变量赋值，SAS 系统把成绩送到 T 变量中，然后转到子程序 COMPUTE 并赋予新值。接着又将改变后的值反馈给原来的变量 t1、t2、t3。

运行程序 3.7 后在输出(OUTPUT)窗口显示出结果，见图 3.7。

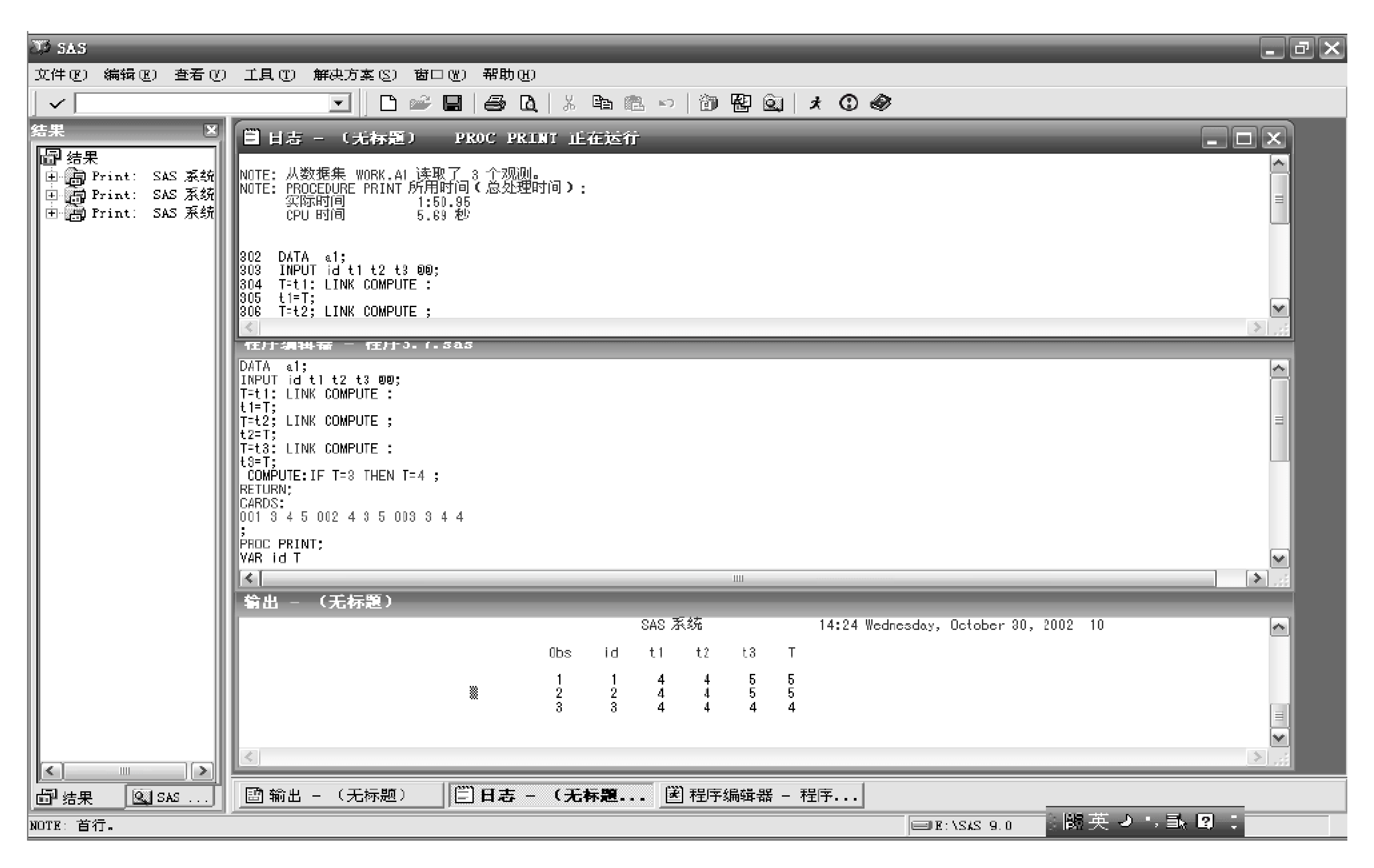


图 3.7 改变 3 次考试成绩

3.4 RETURN 语句

被标识为 OK 的语句对每个个案都执行。但在一定条件下才执行标识时，就应该使用 RETURN 语句。

1. 命令格式

```
RETURN;
```

当是 GOTO…RETURN 结构形式时，遇到 RETURN 语句便返回 DATA 语句并执行 DATA 后面的语句。

当是 LINK…RETURN 结构形式时，遇到 LINK 语句便返回 LINK 语句并执行 LINK 后面的语句。

2. 例子

例 8：GOTO…RETURN 结构，见程序 3.8。

程序 3.8：

```
DATA b1;
INPUT id a b c @@ ;
IF a>b THEN GOTO OK;
  b=4;
  COUNT+1;
```

```
  RETURN;
OK:c+1;
CARDS;
01 5 3 4 02 4 7 8 03 9 7 5 04 6 4 5
;
PROC PRINT;
```

运行程序 3.8 后在输出(OUTPUT)窗口显示出结果,见图 3.8。

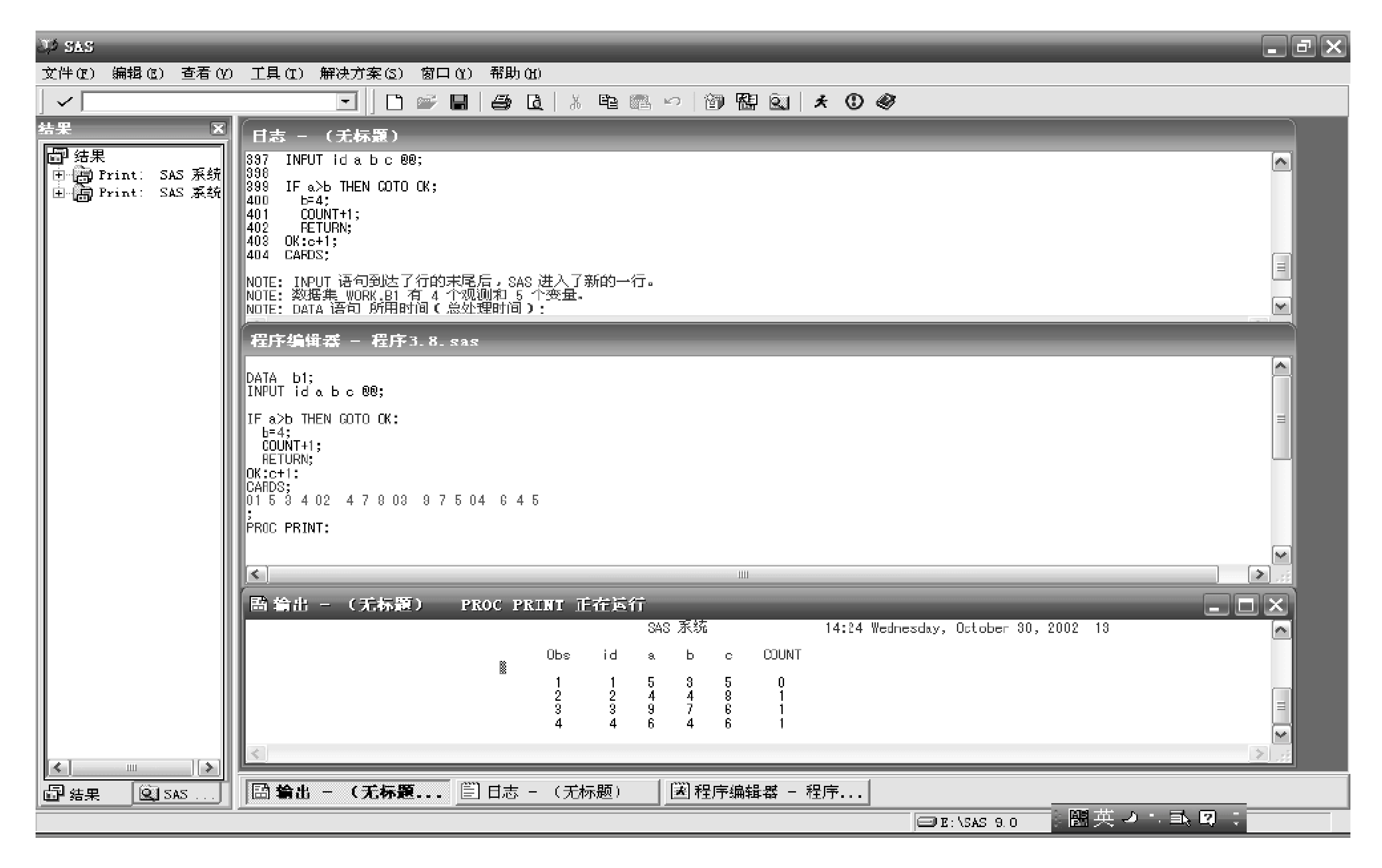


图 3.8 GOTO…RETURN 结构的程序输出

例 9:GOTO…RETURN 结构(见程序 3.9),常用 DO…END 格式替代(见程序 3.10)。

程序 3.9:原来的 GOTO…RETURN 结构语句。

```
DATA G1;
INPUT id a b c @@ ;
IF a>b THEN GOTO OK;
  b=4;
  COUNT+1;
  RETURN;
OK:c+1;
CARDS;
01 5 3 4 02 4 7 8 03 9 7 5 04 6 4 5
;
PROC PRINT;
```

程序 3.10:用 DO…END 格式替代 GOTO…RETURN 结构的语句。

```
DATA G2;/* "GOTO…RETURN"结构常用下面的 DO…END格式替代 */
INPUT id a b @@ ;
IF a>=5 AND a<=8 THEN DO;
    a=6;
  COUNT+1;
END;
ELSE SUMa+b;
CARDS;
001 8.8 5.5 002 7.5 8.5 003 6.5 7.5
;
PROC PRINT;
```

运行程序 3.10 后在输出(OUTPUT)窗口显示出结果,见图 3.9。

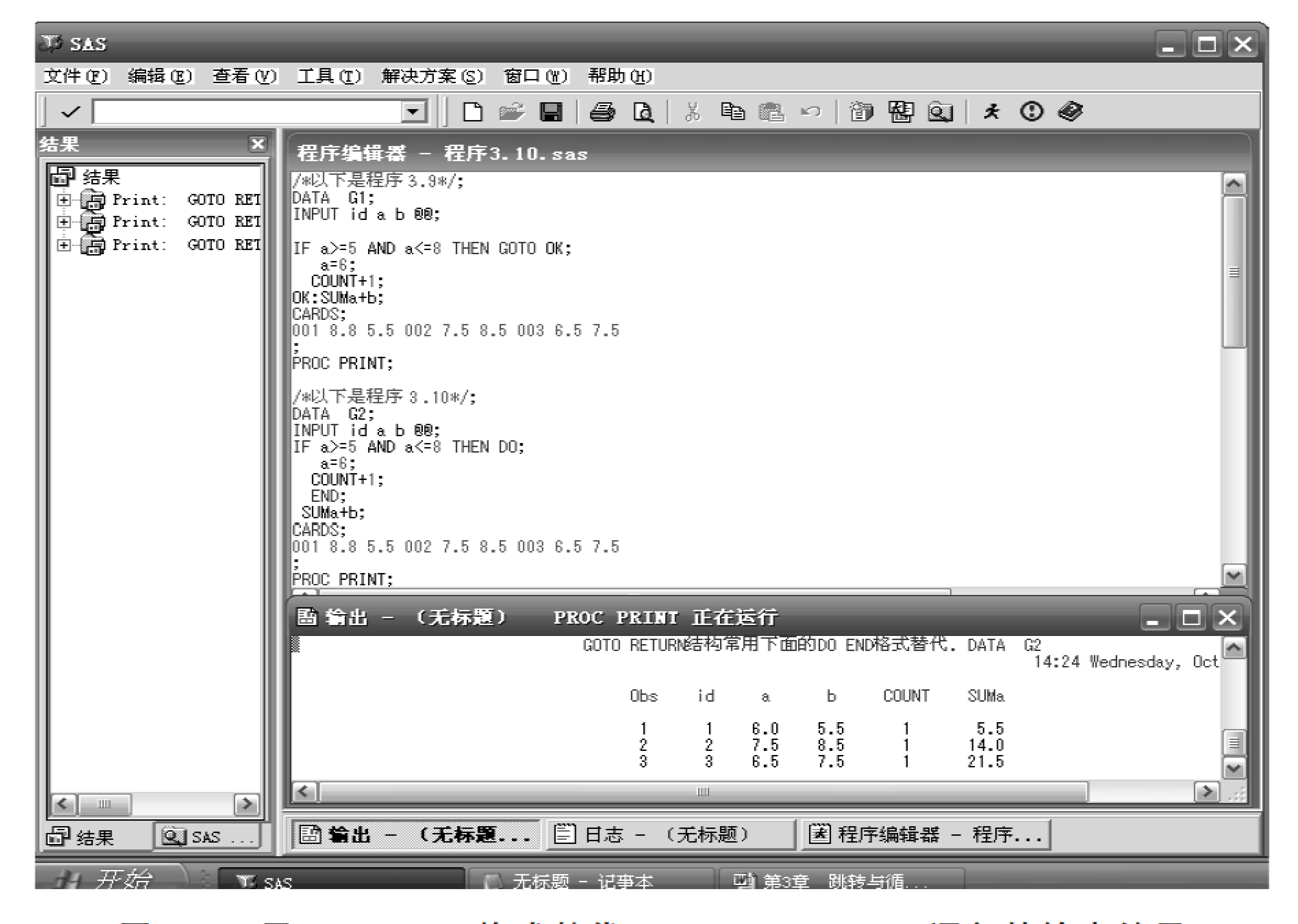


图 3.9 用 DO…END 格式替代 GOTO…RETURN 语句的输出结果

例 10:使用 DO…END 和 IF…THEN/ELSE 两种语句,见程序 3.11。

程序 3.11:

```
/* 程序 3.11:使用 DO…END 和 IF…THEN/ELSE 两种语句 */;
DATA G22;
INPUT id a b @@ ;
IF a>=5 AND a<=8 THEN DO;
    a=6;
  COUNT+1;
  END;
 ELSE SUMa+b;
```

```
CARDS;
001 8.8 5.5 002 7.5 8.5 003 6.5 7.5
;
PROC PRINT;
```

运行程序 3.11 后在输出(OUTPUT)窗口显示出结果,见图 3.10。



图 3.10 使用 DO…END 和 IF…THEN/ELSE 两种语句

说明:程序 3.10 与程序 3.11 的区别在于,前面是用"SUMa+b;"语句,程序 3.11 则用"ELSE SUMa+b;"语句,但结果相同。

3.5 删除部分个案

删除数据集里暂时不用的个案可用 DELETE 语句,或用 IF 语句挑选部分有用的个案数据进行统计。在这种情形下,并不删除原始数据中的个案。

3.5.1 删除数据集里暂时不用的个案

1. DELETE 语句的格式

```
DELETE;
```

2. 例子

例 11:DELETE 语句的使用,见程序 3.12。
程序 3.12:

```
/* 程序 3.12:使用 DELETE 语句 */;
DATA D1;
INPUT id a b c @@ ;
IF a<=65 THEN DELETE;
M=MEAN (OF a b c);
CARDS;
001 88 64 65 002 79 86 90 003 68 77 80
;
PROC PRINT DATA=D1;
```

说明:本例中,变量 a 的值小于或等于 65 分者不进入 SAS 数据集 WORK. D1 中,然后返回到 DATA 语句的下一行继续执行语句。

DELETE 语句常用于 IF…THEN 条件之后作为子句或作为有条件的执行 DO 语句的一部分。

运行程序 3. 11 后在输出(OUTPUT)窗口显示出结果,见图 3. 11 的中部。

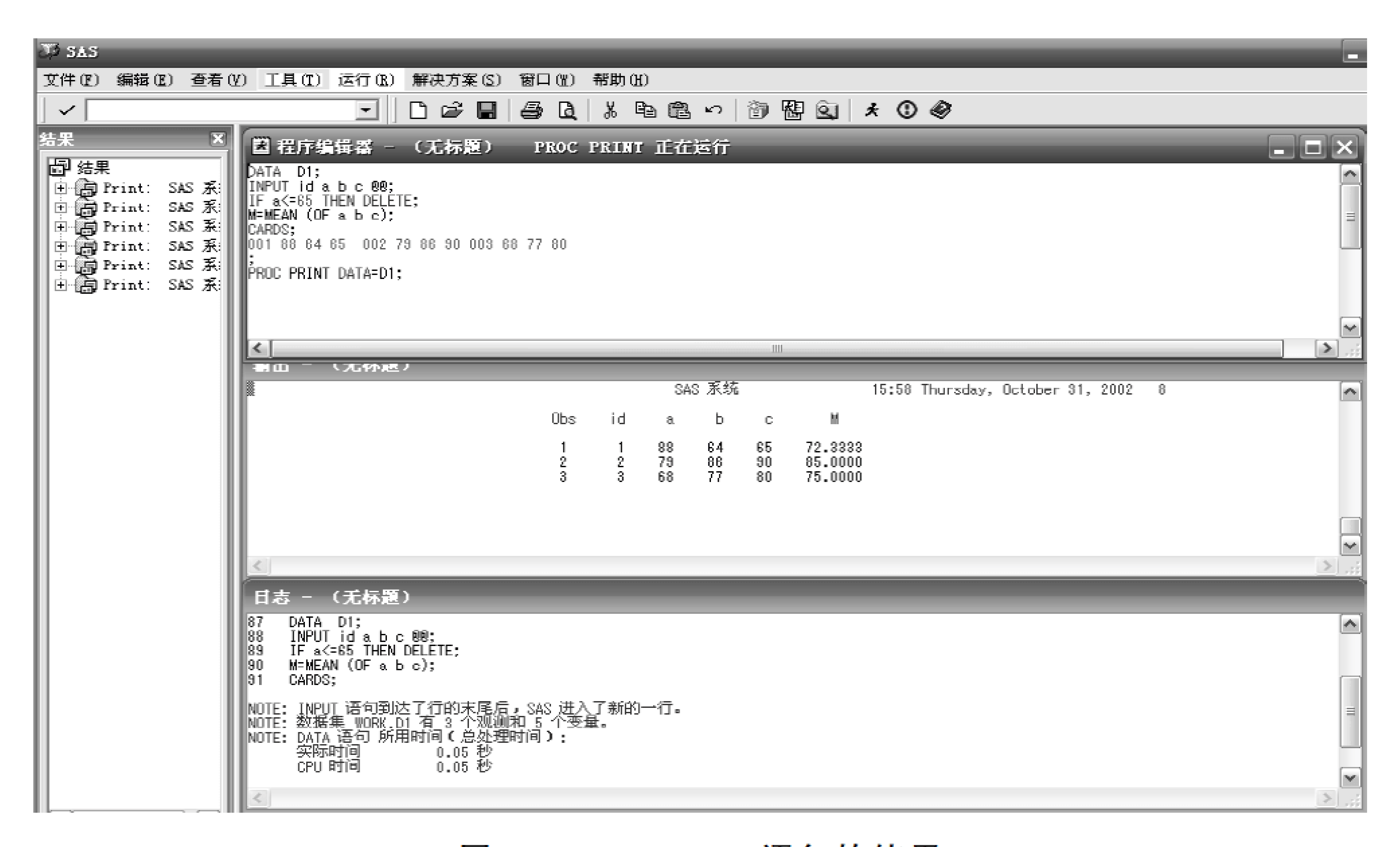


图 3. 11 DELETE 语句的使用

3.5.2 用 IF 语句挑选部分数字型的个案

1. IF 语句的格式

IF <表达式>;

2. 例子

例 12:IF 语句的使用,见程序 3. 13。

程序 3. 13:

```
/* 程序 3.13:使用 IF语句 * /;
DATA I1;
INPUT id sex location age @@ ;
CARDS;
001 1 1 88 002 2 2 79 003 1 2 68 004 2 1 77
;
DATA I2;
  SET I1;
IF sex=1 AND location=1;
PROC PRINT DATA=I2;
DATA I3;
  SET I1;
IF sex=2 AND location=2;
PROC PRINT DATA=I3;
RUN;
```

注意：用 IF 语句挑选数据只能挑选数字型的个案数据，字符型数据不实用！

运行程序 3.13 后在输出(OUTPUT)窗口显示出结果，见图 3.12。

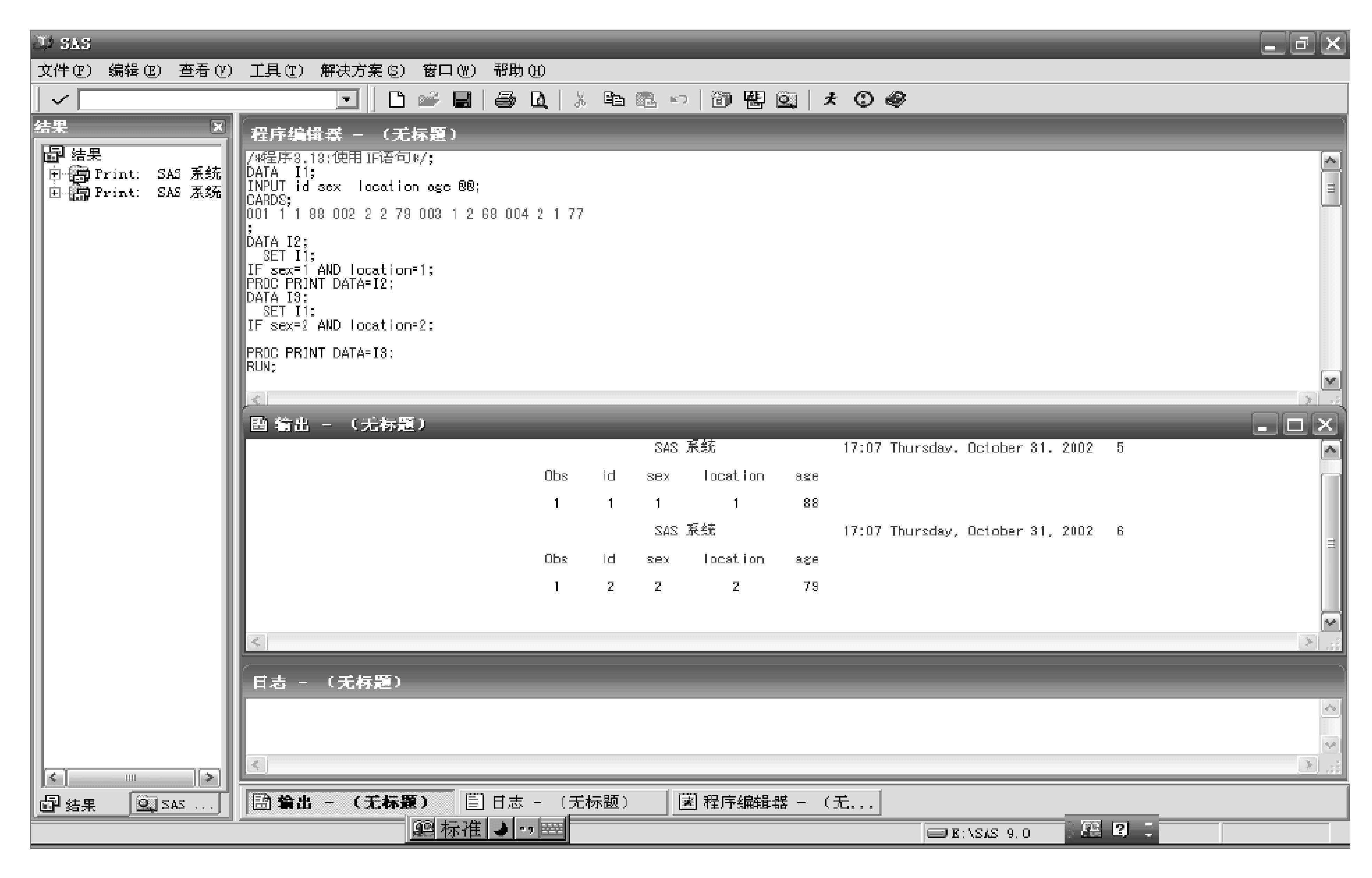


图 3.12 用 IF 语句挖掘部分数字型的个案数据

3.6 循环语句

运行程序时往往要对部分语句(子程序)重复执行多次。例如反复读取某一个值，或反复计数。这就称为循环(语句)。

例 13：由某种条件或指针变量 GOTO，控制反复计数 10 次，见程序 3.14。

程序 3.14：

```
DATA X1;
I=1;
P:PUT I;
I=I+1;
IF I<10 THEN GOTO P;
PROC PRINT DATA=X1;
RUN;
```

运行程序 3.14 后输出结果如图 3.13 所示。

SAS 系统 17:07 Thursday, October 31, 2006

Obs	I
1	10

图 3.13 反复计数 10 次

循环语句可被嵌套在 DO…END 语句之间。

1. DO…END 语句的格式

DO 指针变量=始值 TO 终值 BY 增量;

2. 说明

(1) 始值,终值,增量必须是数字型表达式。

(2) 第 1 次执行 DO 循环语句时,指针变量在始值上。规定了终值和增量后,当执行到 END 语句时,“新的指针变量”=指针变量+增量。然后遇终值进行比较,一旦超过终值,便立即停止执行 DO…END 语句组。

(3) 若未规定始值与终值,则循环组只执行 1 次。若未规定增量,则默认增量为 1。

(4) 第 1 次执行 DO 循环语句之前,就计算始值和终值。

例 14：由指针变量 GOTO 控制反复计数 10 次的完整程序,见程序 3.15。

程序 3.15：

```
DATA X2;
DO A=1 TO 10;
PUT A=;
END;
DATA X3;
DO I=1 BY 1;
PUT I=;
IF I=10 THEN GOTO K;
END;
    K:PUT '循环结束';
PROC PRINT DATA=X3;
RUN;
```

运行程序 3.15 后在“日志(LOG)”窗口显示出结果,见图 3.14。

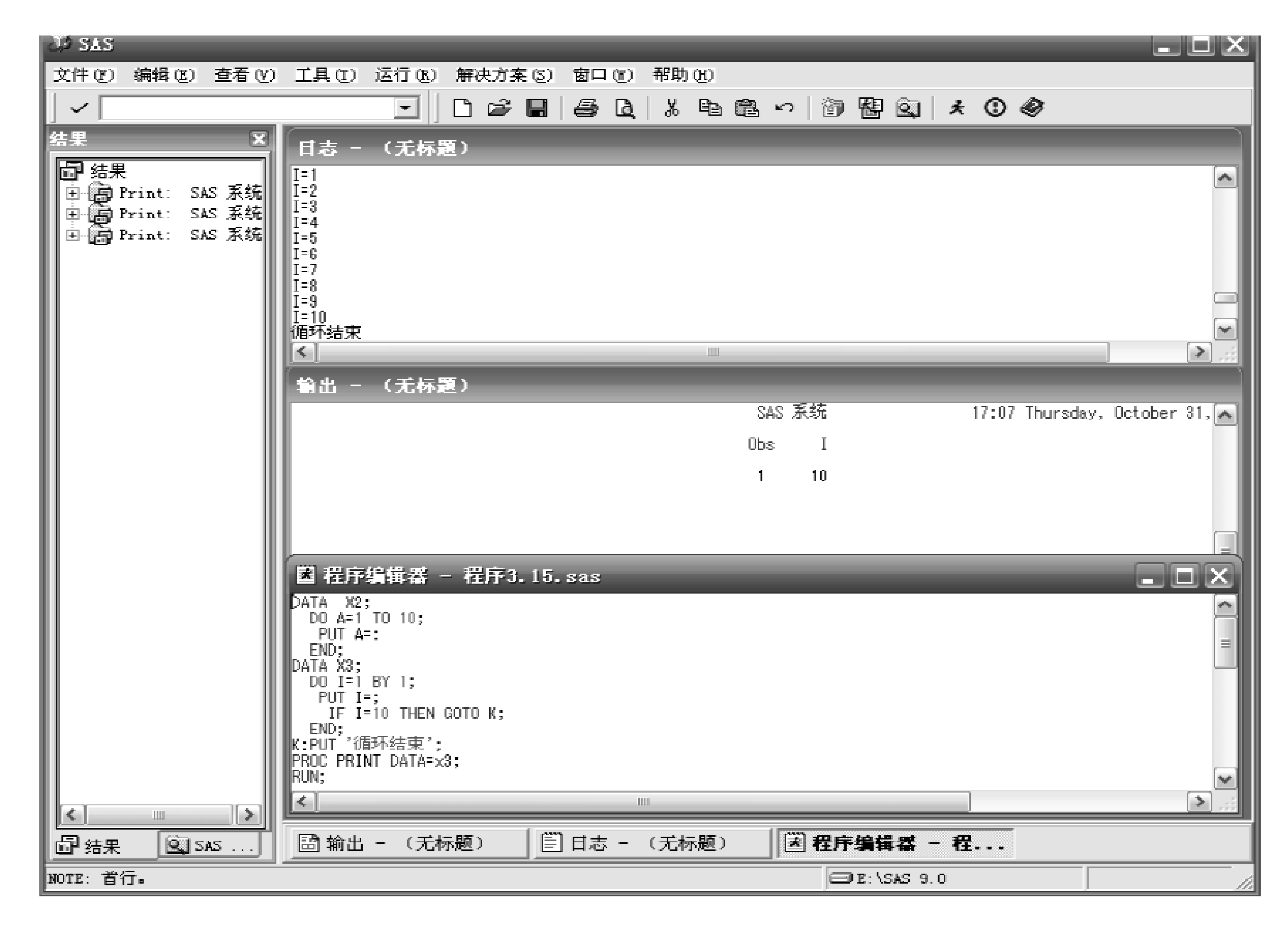


图 3.14 循环 10 次后“循环结束”

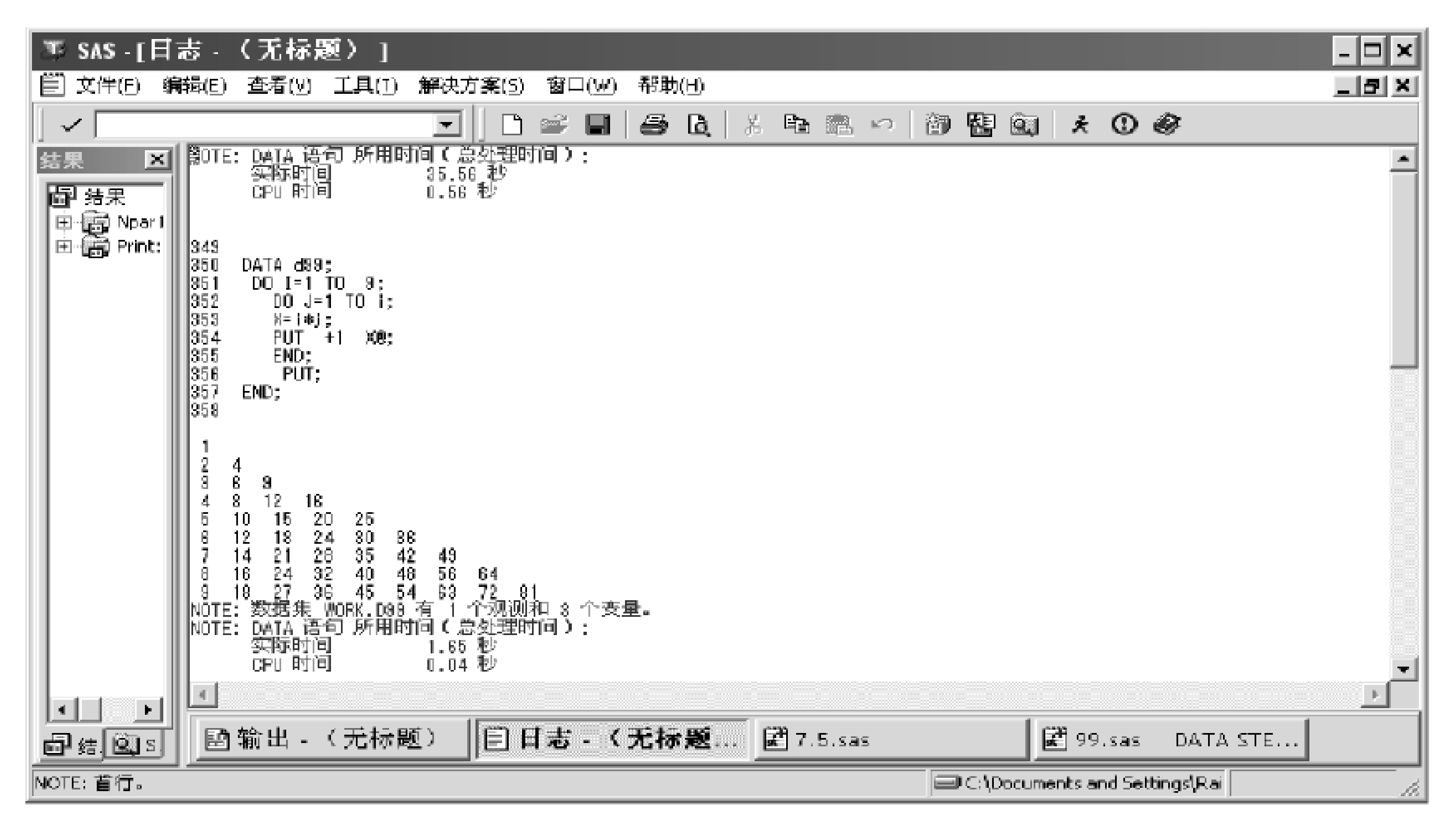
例 15:用循环语句输出“九九”表。

程序 3.16:

```
DATA d9; /* 紧凑的九九表 */
DO I=1 TO 9;
    DO J=1 TO i;
      X=i*j;
      PUT +1 X@ ;
    END;
    PUT;
END;

DATA d99; /* 松散的九九表 */
 DO I=1 TO 9;
   DO J=1 TO I;
      X=i*j;
      PUT +1 @ j +3 X@ ;
   END;
   PUT;
END;
PROC PRINT;
```

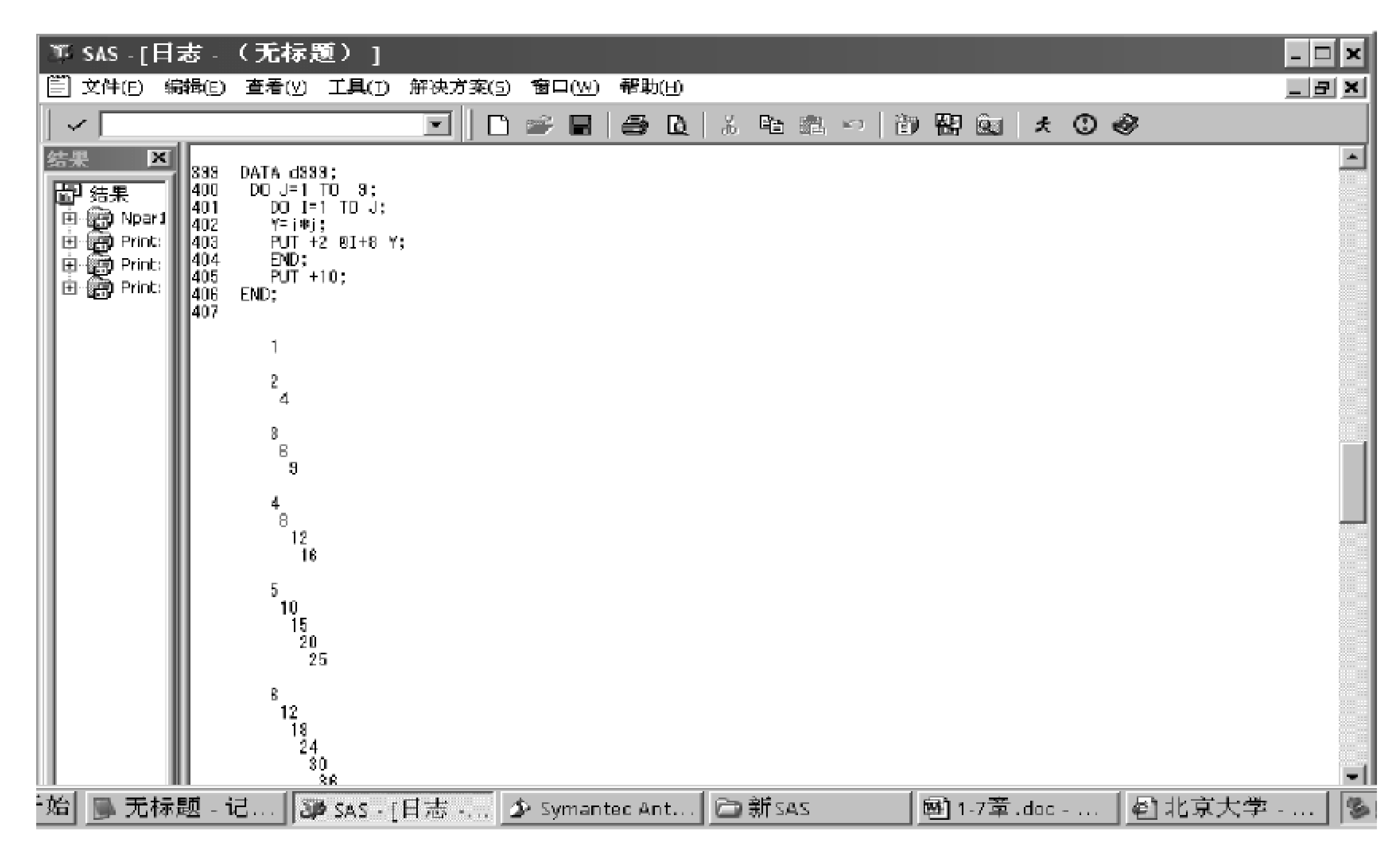
运行程序 3.16 后在“日志(LOG)”窗口显示出结果,见图 3.15(a)和(b)。



(a) 松散的九九表



(b) 紧凑的九九表



(c) 麻花式的九九表

图 3.15　九九表

程序 3.17：麻花式的九九表。

```
DATA d999; /* 麻花式的九九表 */
 DO J=1 TO 9;
   DO I=1 TO J;
   Y=i*j;
   PUT +2 @ I+8 Y;
   END;
   PUT +10;
END;
```

运行程序 3.17 后在“日志(LOG)”窗口显示出结果，见图 3.15(c)。

3.7 数　　组

如果需要以同一方式处理多个变量，则可用 ARRAY 语句把一组变量(数字型或字符型均可)定义为数组中的元素。

1. 语句格式

```
ARRAY 数组名 [n] [$] [长度] 元素 1 元素 2… 元素 n;
```

SAS 是用[n]或{n}指明使用 n 个元素，这个 n 也称为下标。

2. 语句格式说明

[n]或{n}：所引用的数组元素的个数，即下标。

N 为正数时，指定下标的数字，例[n]或{n}表示引用 8 个元素。

N 为 * 号时，清除数组中元素个数的计数，例[*]。

[$]：表示数组中的元素是字符，例如，如果数组中的元素 sex 已在前面定义过字符了(如 INPUT sex $)。则此处的 $ 可省略。

长度：如果在 INPUT 语句中未定义某变量的长度，则用该变量作为元素时必须指定其长度。例如：

```
LENGTH sex $  6;
```

数组元素表：由 INPUT 语句中已定义的变量组成。一个变量可作为不同数组中的元素。但不能把这个数组的元素作为另一个数组中的元素。

3.7.1 下标变量的下标

下标可用数字或任何有效的表达式，并用[ ]或{ }括起来。例如 score [8]表示下标变量是 score，下标元素有 8 个，如 score1-score8。

例 16：提取 score 数组中第 4 个和第 7 个元素。

程序 3.18：

```
DATA s1;
INPUT test1-test5 tt1-tt5@@ ;
ARRAY score{8} test1-test4 tt1-tt4;
    PUT score{4}=    score{7}=;
CARDS;
88 95 80 90 99 100 89 92 89 79
77 68 98 100 88 89 84 78 88 97
;
```

运行程序 3.18 后在“日志(LOG)”窗口显示出 test4 和 tt3 的结果,即第 4 个元素和第 7 个元素,见图 3.16。



图 3.16　挖掘 test4 和 tt3 两个元素

3.7.2 在 DO…END 循环中使用数组

可用 DO 后面的指针变量作为数组的下标。

例 17:将 7 次考试成绩中的 59 分提升为 60 分。

程序 3.19:

```
DATA s2;
INPUT test1-test7 @@ ;
ARRAY score{7} test1-test7;
ARRAY Av7{7} t1-t7;
    DO I=1 TO 7;
  IF score{I}=59 THEN score{I}=60;
    Av7{7}=score{I}/7;
  END;
CARDS;
82 59 77 80 85 88 99 100 59 80 78 86 98 100
```

```
;
PROC PRINT;
```

运行程序 3.19 后在输出(OUTPUT)窗口显示出结果,见图 3.17。

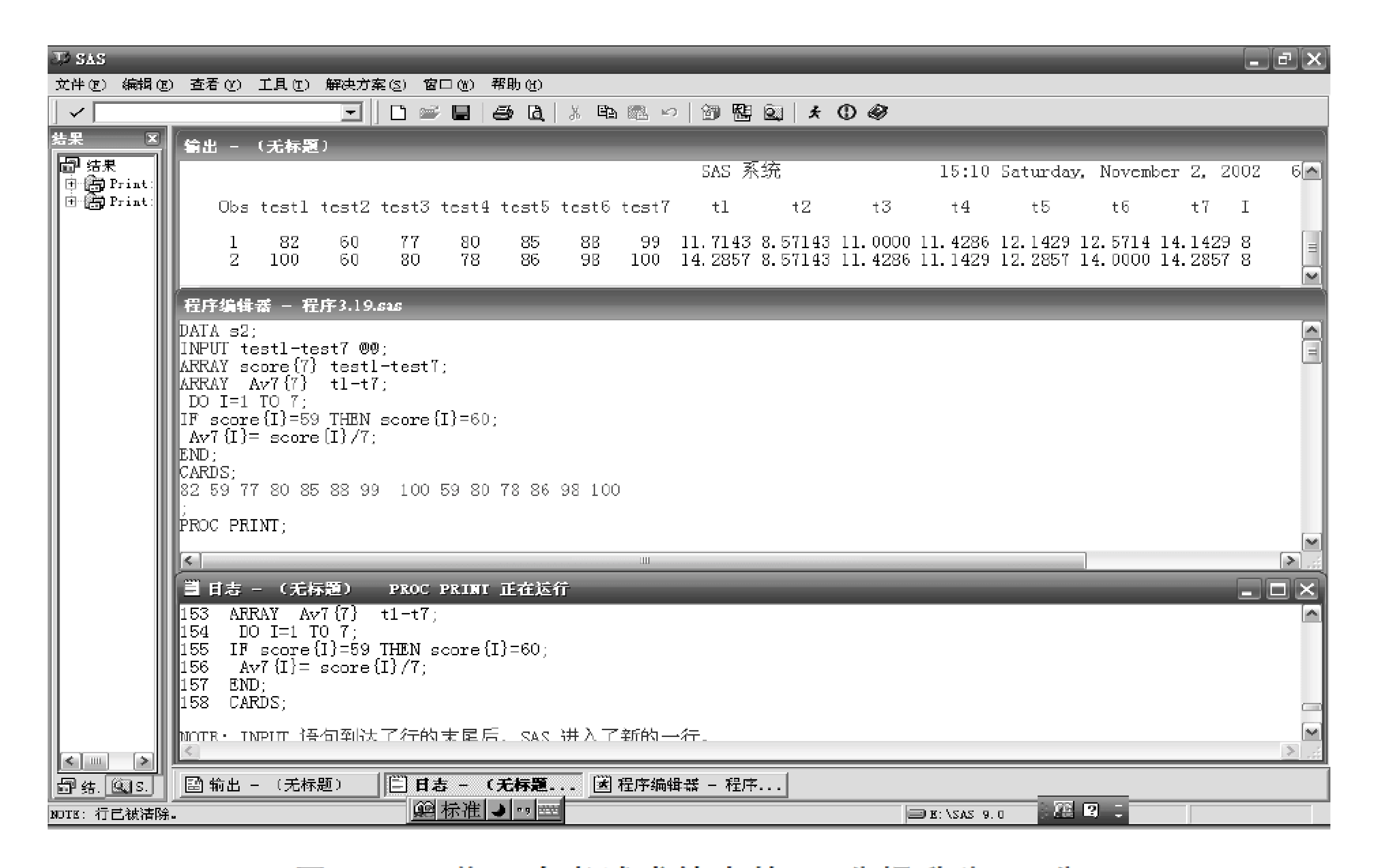


图 3.17 将 7 次考试成绩中的 59 分提升为 60 分

3.7.3 多维数组

有多个下标值的数组是多维数组。

1. 语句格式

```
ARRAY A{4,5} TEST1-TESTn;/* 定义一个 4行 5列的二维数组,用{}或[]均可 */
```

2. 格式说明

A:下标变量名

{4,5}:4 表示 4 行,5 表示 5 列。

TEST1-TESTn:下标变量名可以指定 20 个以内。n<=20。

3. 计算原理

计算机 SAS 系统中,是从数组的左上角开始,将各个变量置于多维数组中。然后按顺序填满各行。例如:

```
ARRAY A{2,3} TEST1-TEST6;
```

此语句将 TEST1 放在{1,1}中,将 TEST2 放在{1,2}中,将 TEST3 放在 TEST3{1,3}中。然后将 TEST4 放在 TEST4{2,1}中,将 TEST5 放在 TEST5{2,2}中,将 TEST6

放在 TEST6{2,3}中。

4. 举例

在恋爱观的各题各项调研中,需要了解每个人对各题各项的选择情况。若有 4 个问题,每个问题有 4 种选择,则是多选项问题。那么,每人就有下面的选择可能:

个案号 Id	变量 v1	v2	v3	v4
01	1	2	3	4
02	4	1	2	3
03	3	1	2	4
04	2	3	3	1

在统计分析时,如果希望输出图 3.18 所示的结果,则应编辑出程序 3.20 所示的命令语句。



Obs	id	v1	v2	v3	v4	d11	d12	d13	d14	d21	d22	d23	d24	d31	d32	d33	d34	d41	d42	d43	d44	I	J
1	1	1	2	3	4	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	5	5
2	2	4	1	2	3	0	0	0	1	1	0	0	0	0	1	0	0	0	0	1	0	5	5
3	3	3	1	2	4	0	0	1	0	1	0	0	0	0	1	0	0	0	0	0	1	5	5
4	4	2	3	3	1	0	1	0	0	0	0	1	0	0	0	1	0	1	0	0	0	5	5

图 3.18 多选项问题的输出结果

程序 3.20:产生图 3.18 的命令语句。

```
DATA p33;
INPUT id v1-v4 @@ ;
ARRAY vv[4] v1-v4;
ARRAY dd[4,4] d11 d12 d13 d14 d21-d24 d31-d34 d41-d44;
  DO I=1 TO 4;
    DO J=1 TO 4;
          IF vv[i]=j THEN DD[i,j]=1;
      ELSE DD[i,j]=0;
    END;
END;
CARDS;
01 1 2 3 4 02 4 1 2 3 03 3 1 2 4 04 2 3 3 1
;
PROC PRINT;
RUN;
```

运行程序 3.20 后输出图 3.18 所示的结果。

习 题 3

1. 用于跳转的语句有哪几种?
2. 试举一个 IF <表达式> THEN…ELSE 的例子。
3. 试举一个 DO…END 的例子。
4. 试举一个 GO TO 语句的例子。
5. 试用循环语句输出“九九”表。

第 4 章 chapter 4

建立数据集

数据挖掘前需要有数据库(或称数据仓库),按 SAS 的行话讲,就是建立 SAS 数据集。

4.1 建立永久数据集

SAS 系统一经启动就在硬盘中开辟出一个临时的工作区,称为 WORK.*,并把 DATA 步(DATA 语句)产生的数据存储为临时数据集 WORK.DATA1,或产生人为指定的数据集,如 WORK.MY1。但是一旦退出 SAS 系统或关机,临时数据集便丢失。

为了将这类临时数据集存储为永久性的数据集,就要指定存储数据集的盘符及子目录名(即路径)和文件名,通常是通过 LIBNAME 语句及 DATA 语句实现的。

1. 先用 LIBNAME 语句指定路径

SAS 的数据集名称前面是无法指定真实的路径的,必须先用 LIBNAME 语句给路径起个别名(即“库逻辑名”),以后就可在 DATA 语句中将别名还原为真实的路径名。

1) LIBNAME 语句格式

```
LIBNAME语句 别名 '路径名';
```

2) 格式说明

LIBNAME 语句:一旦出现一个 LIBNAME 语句,便一直起纽带的联系作用,直到重新指定另一个 LIBNAME 语句才改变为另外的路径联系。

别名:库逻辑名。如 LA、L 或其他字母。

路径名:希望存入永久数据集的磁盘及其子目录,并且用一对左撇号(或左引号)括起来。如'D:\SAS\MY1'或'D:\MY1'。

3) 例子

例 1:将数据集永久存储于'D:\SAS 9\MY1'中。其命令语句如下。

```
LIBNAME LB 'D:\SAS 9'; /* 在 D盘建立一个库逻辑名为 LB,然后把 D:\SAS 9
                          这个路径赋予 LB* /
```

2. 再用 DATA 语句确认 LB 名为实名

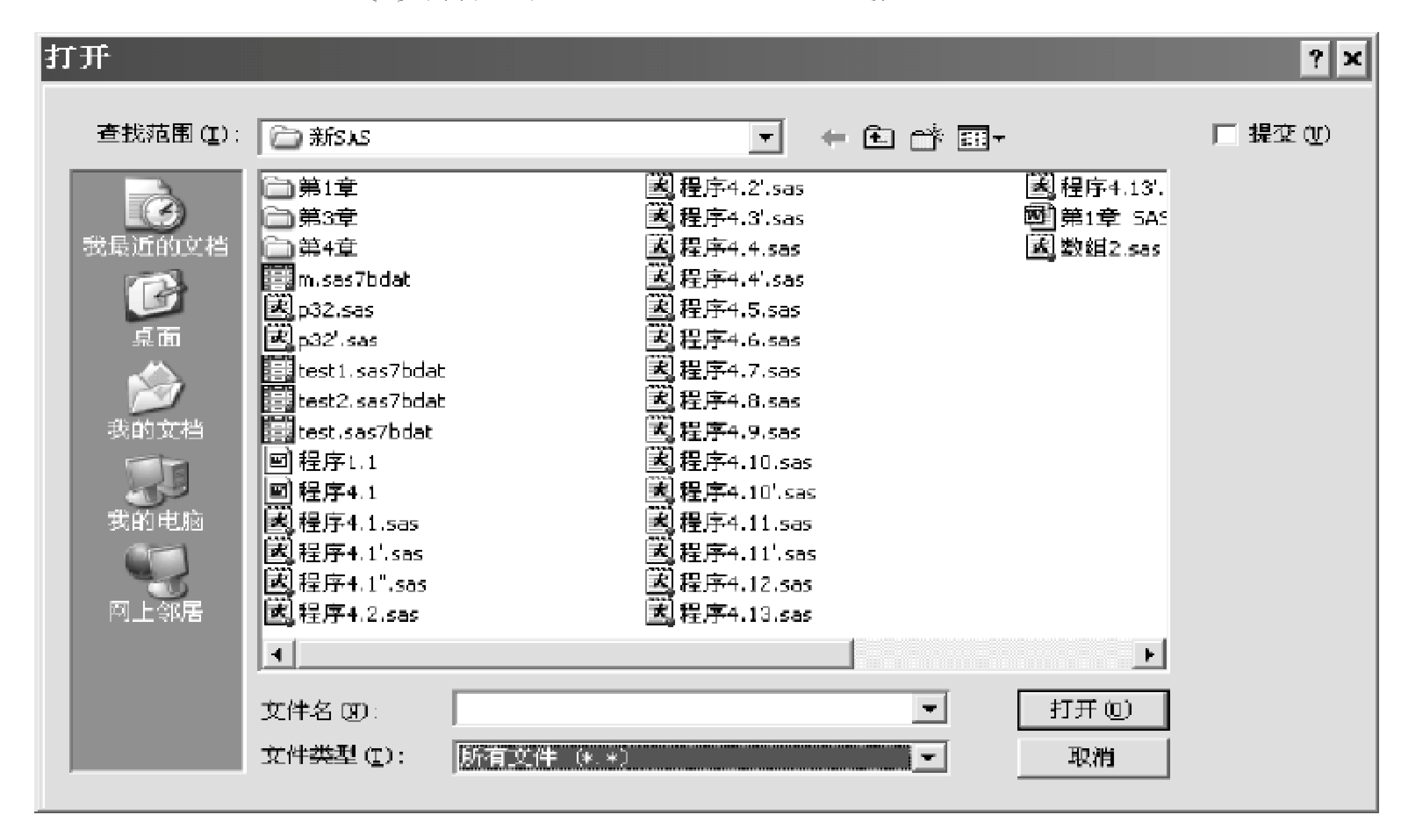
```
DATA LB.MY1;          /* 在 DATA语句中确认 LB名并指定文件名 MY1就可
```

建立一个永久的数据集“D:\SAS 9\MY1.SAS7BDAT”。
“.SAS7BDAT”是计算机 SAS 内部自动加上去的默认后缀 * /

例 2：将数据集 m.SAS7BDAT 永久存储于'D:\新 SAS'中，见图 4.1(a)。其命令语句见程序 4.1a。



(a) 数据集 m.SAS7BDAT 的信息



(b) 存储数据集 D:\新 SAS\m.SAS7BDAT

图 4.1　将数据集 m.SAS7BDAT 永久存储于'D:\新 SAS'中

程序 4.1a：

```
LIBNAME LB 'D:\新 SAS';
```

```
DATA LB.m;
INPUT id sex $  age h w @@ ;
CARDS;
01 m 20 1.68 65.5 02 f 25 1.70 68.5
03 m 28 1.71 69.5 04 f 30 1.69 59.5
;
PROC PRINT;
RUN;
```

运行程序 4.1a 产生图 4.1 所示的结果。

例 3：同一个 DATA 步(程序)中可定义多个 LIBNAME 语句。例如，在第 1 个路径中读出永久的 SAS 数据集 TEST1. SAS7BDAT 数据集，在第 2 个路径中读出永久的 SAS 数据集 TEST2. SAS7BDAT 数据集，见程序 4.1b。

程序 4.1b：

```
LIBNAME LB2 'F:\新 SAS';     /* 定义 F:\新 SAS 路径 */
LIBNAME LB3 'D:\新 SAS';     /* 再定义 D:\新 SAS 路径 */
DATA LB2.test2;              /* 建立永久的数据集 F:\新 SAS\test2.SAS7BDAT */
INPUT id sex $  age h w;
CARDS;
01 m 20 1.68 65.5
02 f 25 1.70 68.5
03 m 28 1.71 69.5
04 f 30 1.69 59.5
;
PROC PRINT DATA=LB2.test2;   /* 见图 4.2(a)倒数第 2 集的数据 */
DATA LB3.test3;              /* 建立永久的数据集 D:\新 SAS\test3.SAS7BDAT */
SET LB2.test2;               /* 把 LB2.test2 数据集复制给 LB3.test3 */
KEEP sex age;                /* 永久数据集 D:\新 SAS\test3.SAS7BDAT 中只保留 sex 和 age 变量 */
PROC PRINT DATA=LB3.test3;   /* 见图 4.2(a)最后一个数据集的数据 */
RUN;
```

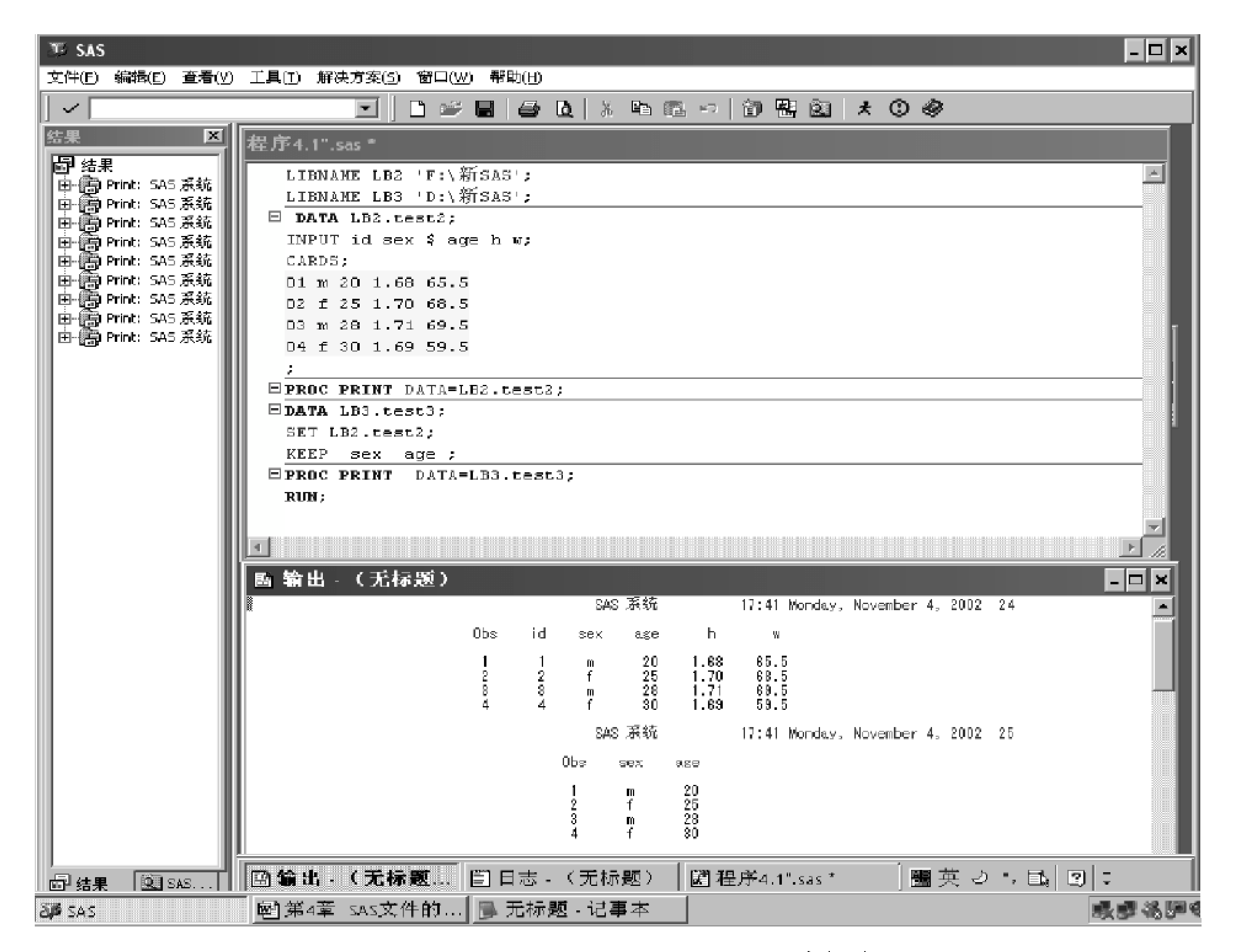
运行程序 4.1b 产生图 4.2(b)所示的结果。

一旦建立了永久的 SAS 数据集(如图 4.2(a)的 m. SAS7BDAT 或图 4.2(b)test2. SAS7BDAT)，以后只要从该路径(如 F:\新 SAS\)调出数据集(test2. SAS7BDAT)，便可进行统计分析。

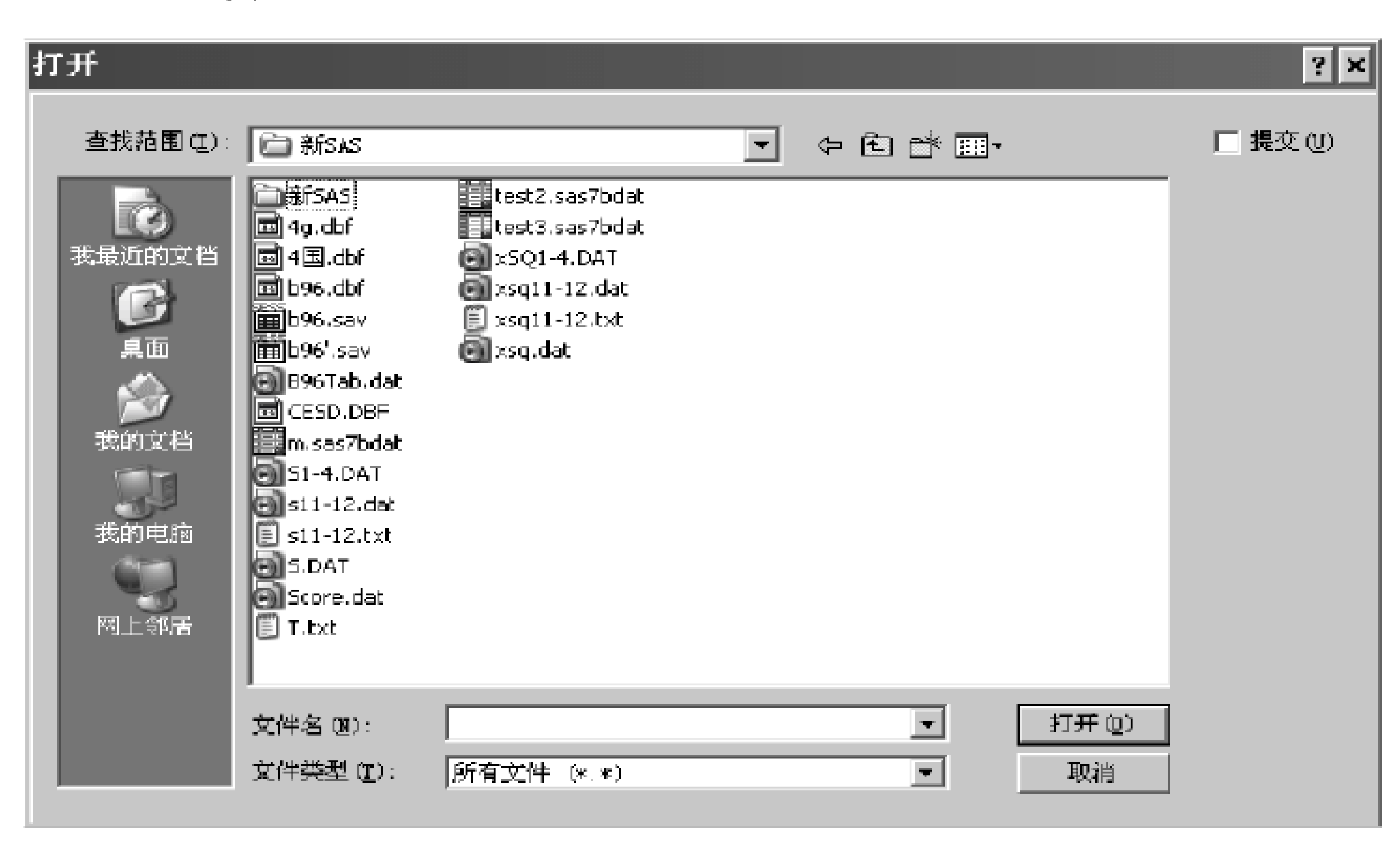
例 4：直接调用 F:\新 SAS\test2. SAS7BDAT 数据集进行显示，见程序 4.2。

程序 4.2：

```
LIBNAME LB3 'D:\新 SAS';      /* 先指定要显示的数据集路径名称"D:\新 SAS 路径" */
PROC PRINT DATA=LB3.TEST3;    /* 然后显示数据集"D:\新 SAS\test3.SAS7BDAT" */
LIBNAME LB2 'F:\新 SAS';      /* 先指定要显示的数据集路径名称"F:\新 SAS 路径" */
PROC PRINT DATA=LB2.TEST2;    /* 然后显示数据集"F:\新 SAS\test2.SAS7BDAT" */
```



(a) 数据集 F:\新 SAS\test2.SAS7BDAT 的信息



(b) 存储数据集 F:\新 SAS\test2.SAS7BDAT

图 4.2　输出 2 个数据集的信息

运行程序 4.2 产生图 4.3 所示的结果。

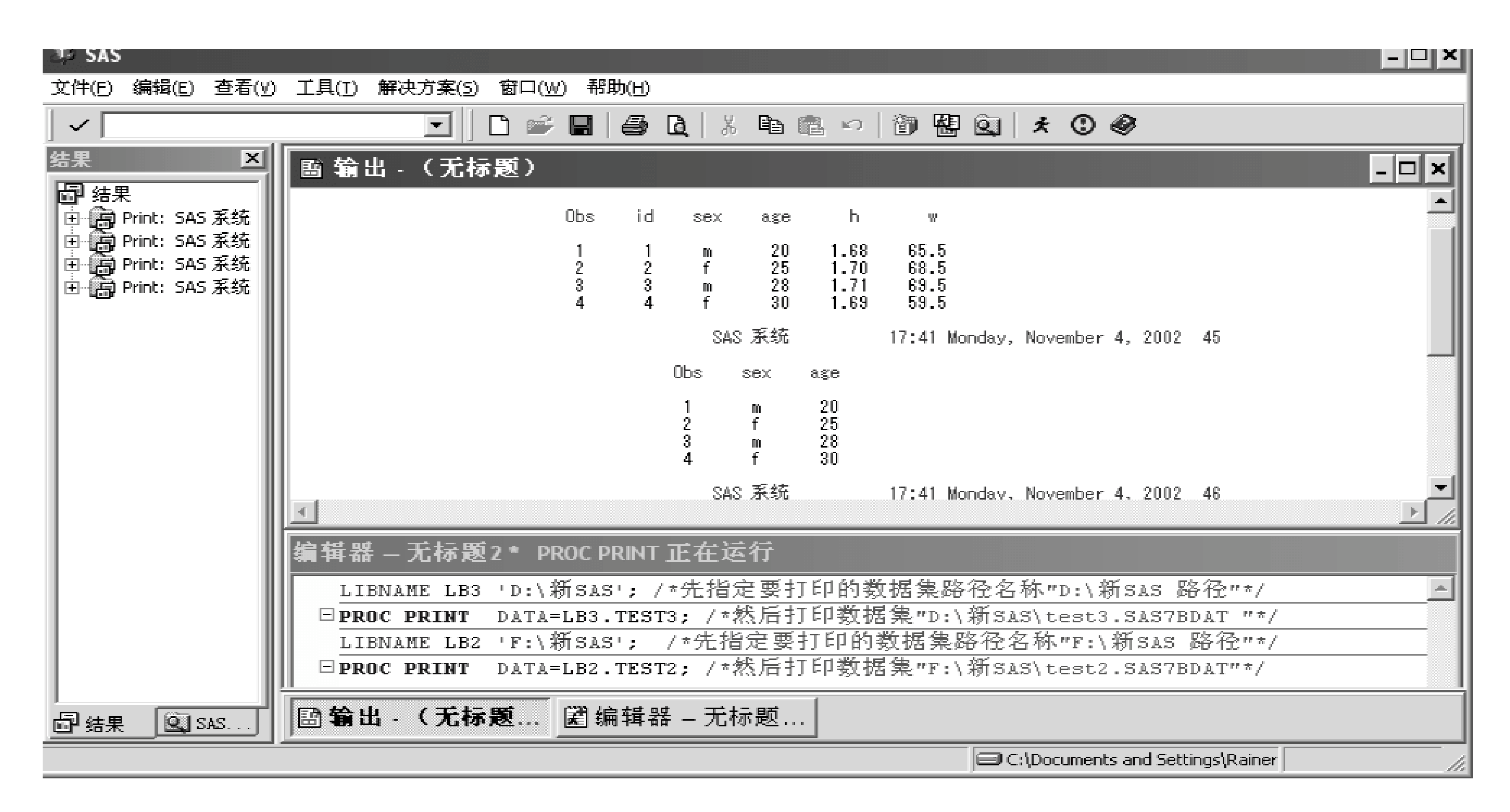


图 4.3 输出数据集 D:\新 SAS\test3.SAS7BDAT 的信息

4.2 数据的分组及分组标记

数据的分组是为了按照某个变量值(如性别 sex 的男女两组)排序数据,以便分组统计分析。

4.2.1 分组控制

分组控制的语句为 BY 语句。BY 语句总是和 SORT 语句联合使用,以便按组预先排序个案后再执行 SET(复制)、MERGE(合并)、UPDATA(更换数据)等操作。

1. BY 语句格式

```
BY [DESCENDING] 变量 1 变量 2…… ;
```

2. 说明

(1) 默认为按照变量 1 的值对个案升序排序。

(2) [DESCENDING] 变量 1:按照变量 1 的值对个案降序排序,其值相同时再按照后面的变量(如“变量 2”)升序排序个案。

例 5:按性别升序排序个案。

程序 4.3a:

```
DATA test1;
INPUT id sex $  age h w @@ ;
CARDS;
01 m 20 1.68 65.5 02 f 25 1.70 68.5
```

```
03 m 28 1.71 69.5 04 f 30 1.69 59.5
;
DATA test2;
SET test1;
PROC SORT;
BY sex age h w ;
PROC PRINT;
RUN;
```

运行程序 4.3a 产生图 4.4 所示的结果。



图 4.4 按性别升序排序个案

例 6：先按 sex 值升序排序个案，性别相同者再按年龄降序排序个案。
程序 4.3b：

```
DATA test1;
INPUT id sex $  age h w @@ ;
CARDS;
01 m 20 1.68 65.5 02 f 25 1.70 68.5
03 m 28 1.71 69.5 04 f 30 1.69 59.5
;
DATA test2;
SET test1;
PROC SORT;
BY sex DESCENDING age h w ;
PROC PRINT;
RUN;
```

运行程序 4.3b 产生图 4.5 所示的结果。

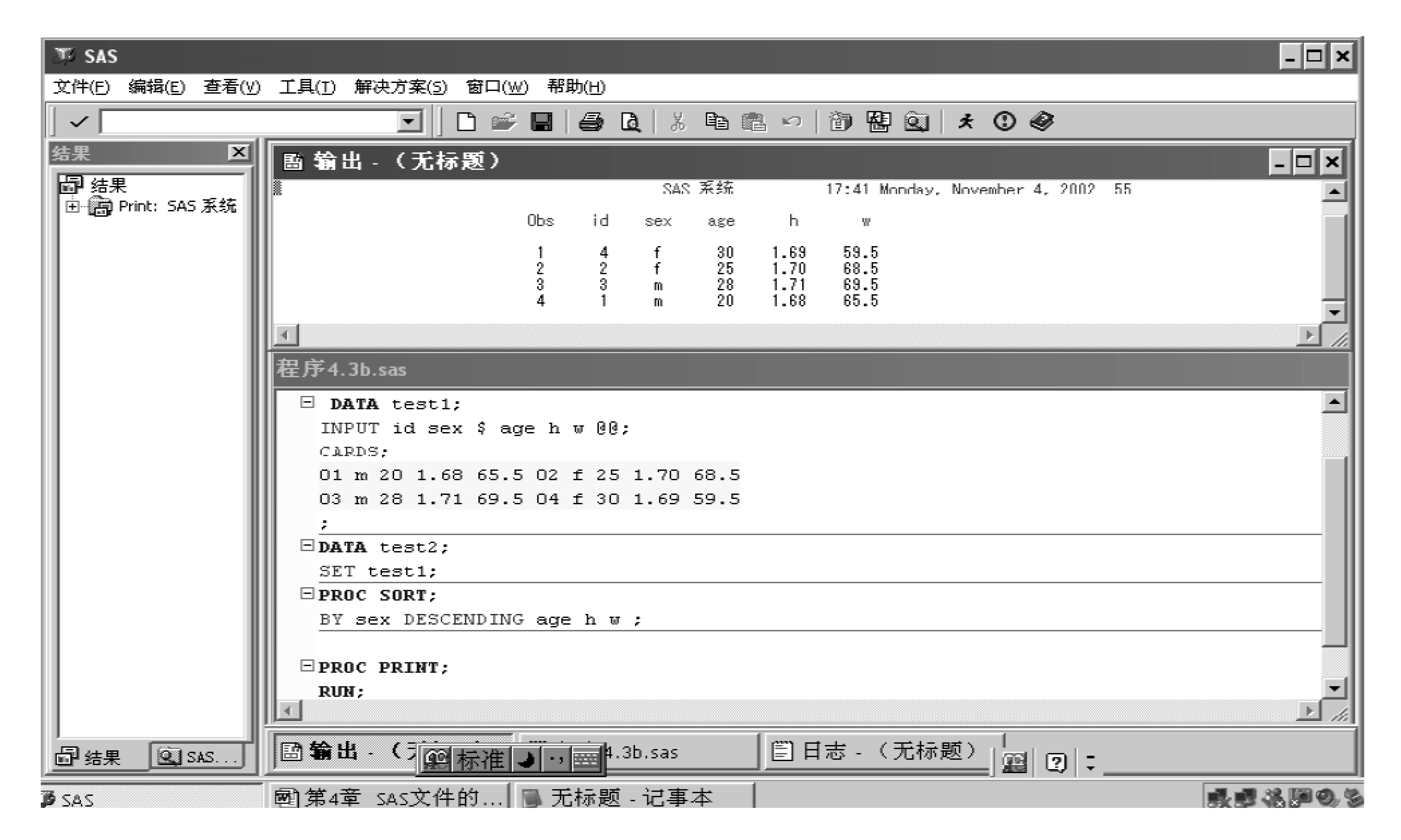


图 4.5 按 sex 值升序排序个案，性别相同者再按年龄降序排序个案

4.2.2 数据的分组标记

用 BY 语句分组处理数据集时(如 BY sex 之类的语句)，将建立两个变量：FIRST. sex 和 LAST. sex，以便对分组时的首尾两个个案做记录。

当检测到某个个案是按 sex 值分组的第一个人(或最小值)时，则 FIRST. sex=1；否则 FIRST. sex=0。

当检测到某个个案是按 sex 值分组的最后一个人(或最大值)时，则 LAST. sex=1；否则 LAST. sex=0。

FIRST. sex 和 LAST. sex 只用于 DATA 语句中进行编程，但不保存到数据集里。

例 7：程序 4.4。

```
LIBNAME LBa 'F:\新 SAS';
 DATA LBa.test;
 INPUT id location $  sex $  t1-t3 @@ ;
 TOTAL= SUM(OF t1-t3);
 CARDS;
 01 D m 78 88 90 02 X F 88 76 92
 03 N m 90 78 88 04 B f 69 84 99
 ;
 PROC PRINT;
 PROC SORT DATA= LBa.TEST OUT= score1;
 BY location sex;
 DATA score2;
 SET score1;
```

```
BY location sex;
DROP T1-T3;
IF FIRST.LOCATION THEN PUT id location sex;
PROC PRINT ;
VAR id location sex;
RUN;
```

运行程序 4.4 产生图 4.6 所示的结果。

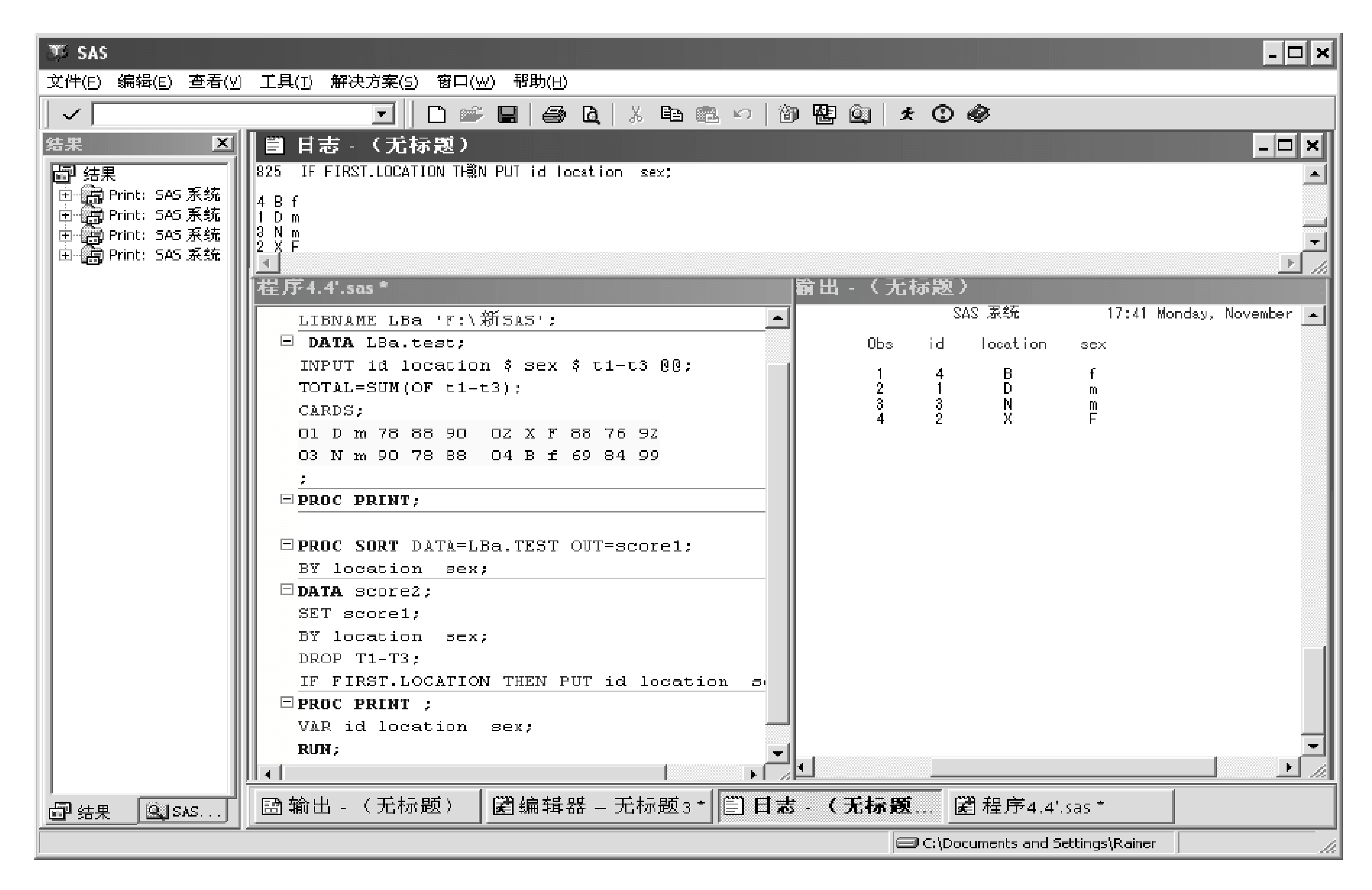


图 4.6 在日志窗口输出 FIRST. LOCATION 值

从图 4.6 的日志窗口显示出 FIRST. LOCATION＝1 时的 4 行观测值，同时右侧输出窗口详细显示 4 个观测值的数据。

4.3 数据的排序

数据排序是统计分析必经的过程。例如对收入、年龄或成绩进行升序排序等。

1. 语句格式

```
PROC SORT DATA=数据集名 OUT=输出名;
  BY [DESCENDING] 变量 1 变量 2 ;
```

2. 例子

例 8：按地区、性别对学生成绩的数据集“F:\新 SAS\test. SAS7BDAT”进行排序，见程序 4.5。

程序 4.5：

```
LIBNAME LB1 'F:\新 SAS';
DATA LB1.test;
INPUT id location $  sex $  t1-t3 @@ ;
TOTAL=SUM(OF t1-t3);
CARDS;
01 D m 78 88 90 02 X F 88 76 92
03 N m 90 78 88 04 B f 69 84 99
05 N f 80 88 90
;
PROC PRINT; /* 显示数据集里的数据,但尚未排序 */
PROC SORT DATA=LB1.test OUT=score1;
BY location sex;
DATA score2;
SET score1;
BY location sex;
IF FIRST.LOCATION THEN PUT id location sex;
PROC PRINT; /* 按地区升序排序并显示数据集里的数据 */
RUN;
```

运行程序 4.5 产生图 4.7 所示的结果。

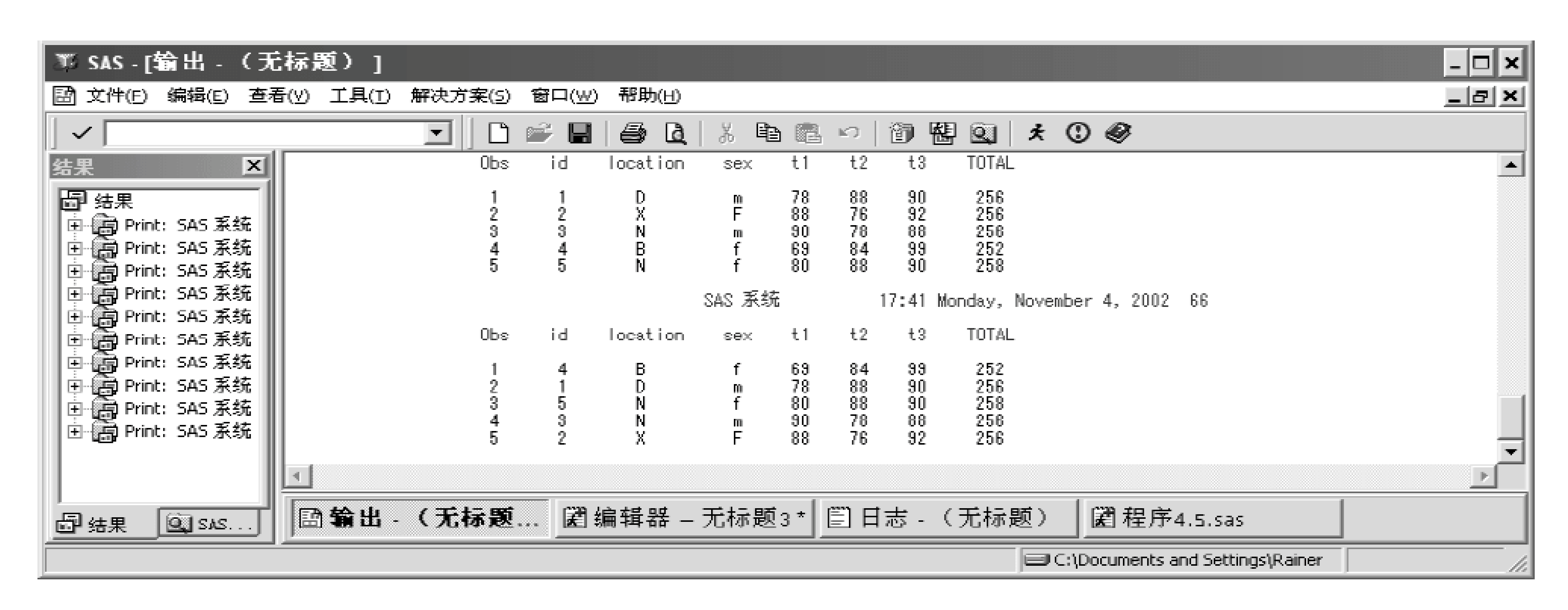


图 4.7 未排序和排序的结果对比

4.4 数据集的连接

可用 SET 语句将几个数据集一个接一个地连接成大的数据集。第 2 个数据集连接在第 1 个数据集的后面。

4.4.1 变量相同时的连接

变量相同时的连接是纵向地将几个数据集的个案连接起来。变量的个数不变但个案增多了。

例如,数据集 a 中有 300 个个案,每人有 id、sex、age 三个变量。数据集 b 中有 500 个

个案，每人也有 id、sex、age 三个变量。那么，连接成为大的数据集之后，将有 800 个个案，每个个案仍然是 id、sex、age 三个变量。

例 9：两个数据集的连接，见程序 4.6。

程序 4.6：

```
DATA A;
INPUT id sex $  age income @@ ;
CARDS;
01 m 38 2000 02 f 30 2100
;
DATA B;
INPUT id sex $  age income ;
CARDS;
03 m 31 2050
;
DATA AB1;
SET A B;
PROC PRINT DATA= ab1;
RUN;
```

运行程序 4.6 产生图 4.8 所示的结果。

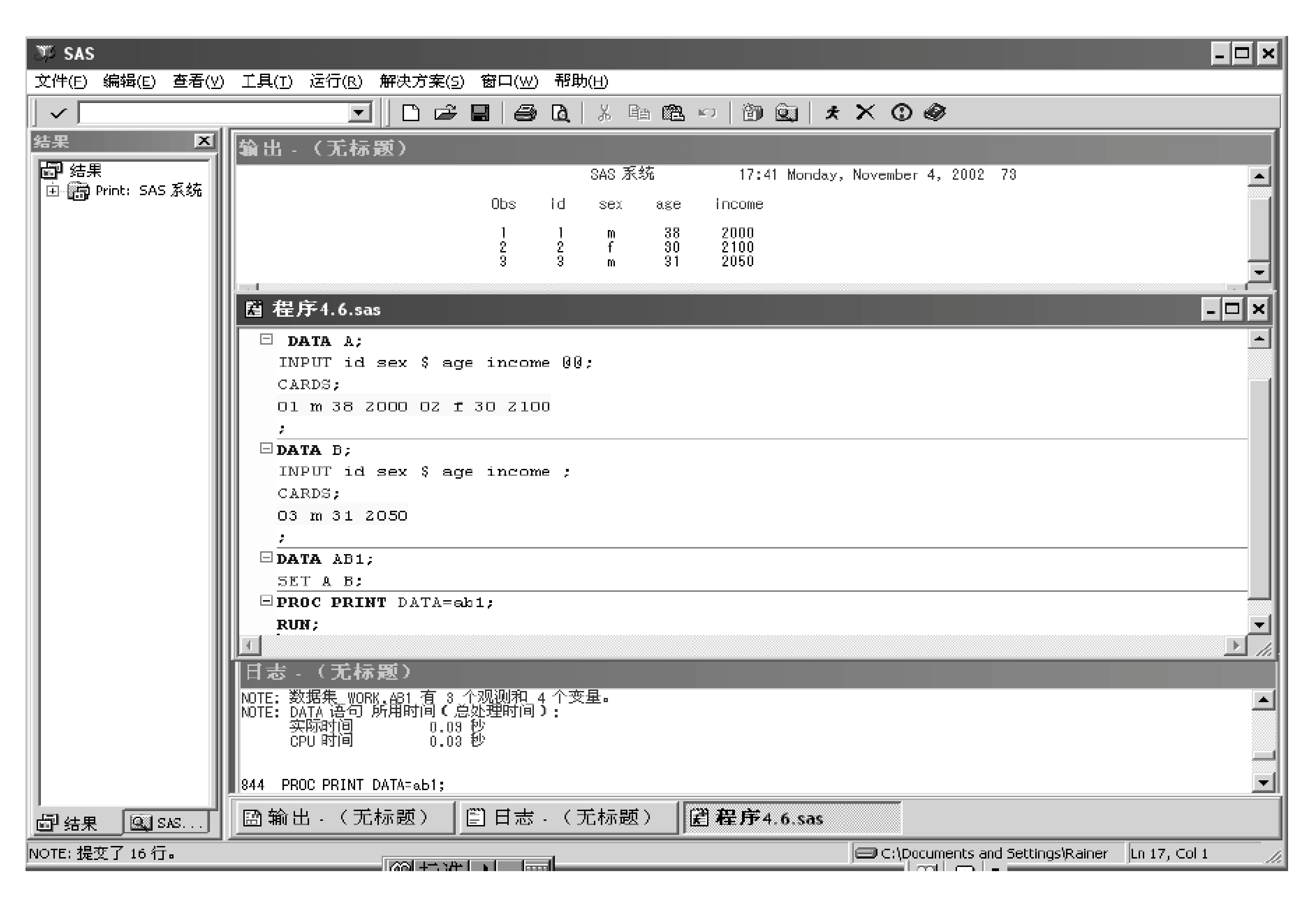


图 4.8　两个数据集的连接

4.4.2　变量不同时的连接

变量不同时的连接是横向地将几个数据集的变量连接起来。变量的个数变多了但

个案数目不变。

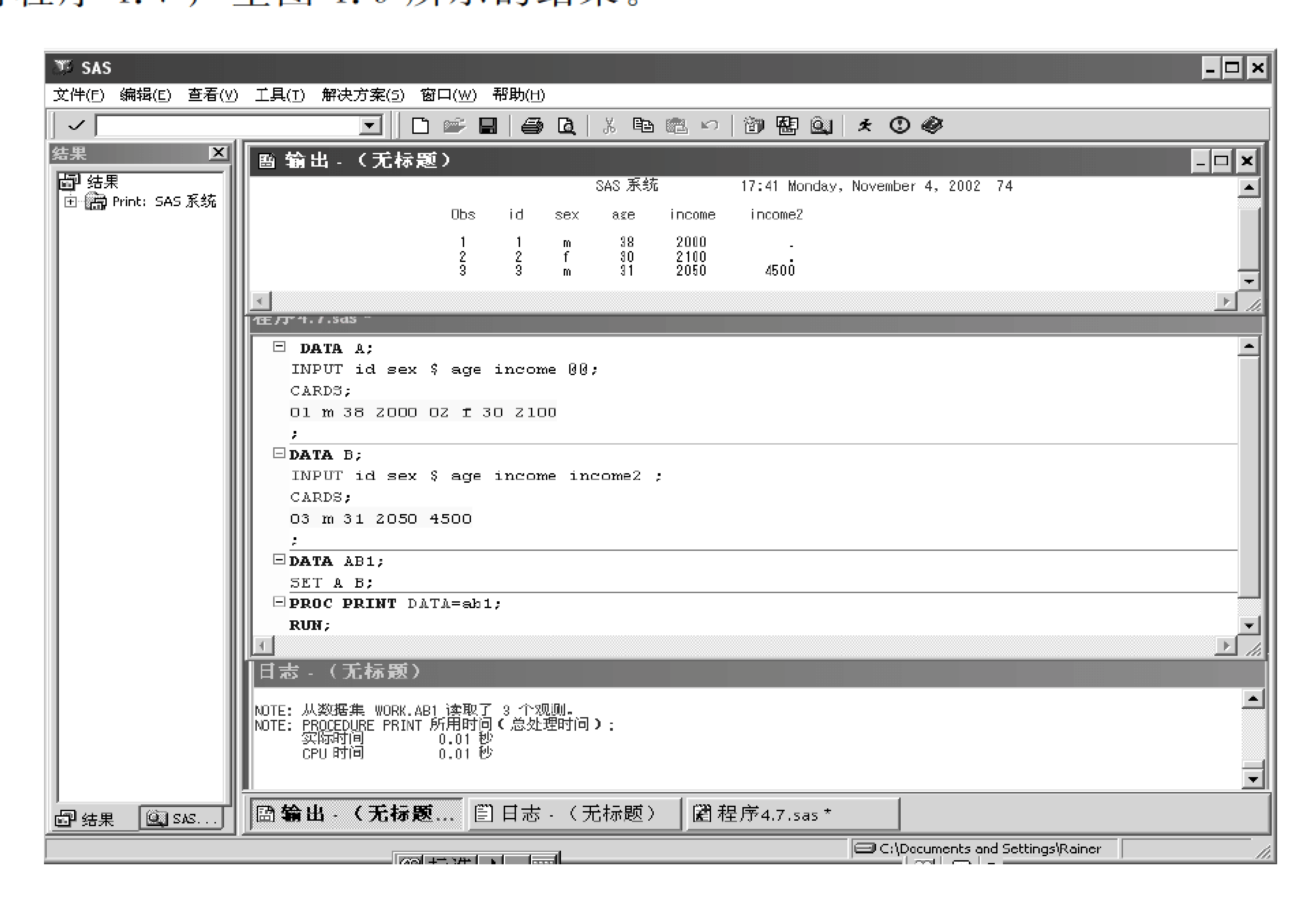
例如，数据集 a 中有 300 个个案，每人有 id、sex、age 三个变量。数据集 b 中有 300 个个案，每人另有 inc1、inc2、inc3 三个变量。那么，连接成为大的数据集之后，仍然有 300 个个案，但是每个个案的变量被连接为 3+3=6 个变量。

例 10：变量不同时的连接，见程序 4.7。

程序 4.7：

```
DATA A;
INPUT id sex $  age income @@ ;
CARDS;
01 m 38 2000 02 f 30 2100
;
DATA B;
INPUT id sex $  age income income2 ;
CARDS;
03 m 31 2050 4500
;
DATA AB1;
SET A B;
PROC PRINT DATA= ab1;
RUN;
```

运行程序 4.7 产生图 4.9 所示的结果。



Obs	id	sex	age	income	income2
1	1	m	38	2000	.
2	2	f	30	2100	.
3	3	m	31	2050	4500

图 4.9　变量不同时的连接

4.4.3 变量值相同时的个案连接

如果两个数据集 A 和 B 是按年龄升序排序的，连接时要求将年龄相同的个案连接在一起，这就是变量值相同时的个案连接。这时要用 BY 语句进行同值的个案连接。

例 11：变量值相同时的个案连接。

程序 4.8：

```
DATA A;
INPUT id sex $  age income @@ ;
CARDS;
01 m 38 2000 02 f 38 2100
;
DATA B;
INPUT id sex $  age income income2;
CARDS;
03 m 31 2050 4500
;
PROC SORT OUT= SORTa;
BY age;
PROC SORT OUT= SORTb;
BY age;
DATA AB3;
SET A B;
PROC SORT;
   BY age;
PROC PRINT DATA= ab3;
RUN;
```

运行程序 4.8 产生图 4.10 所示的结果。

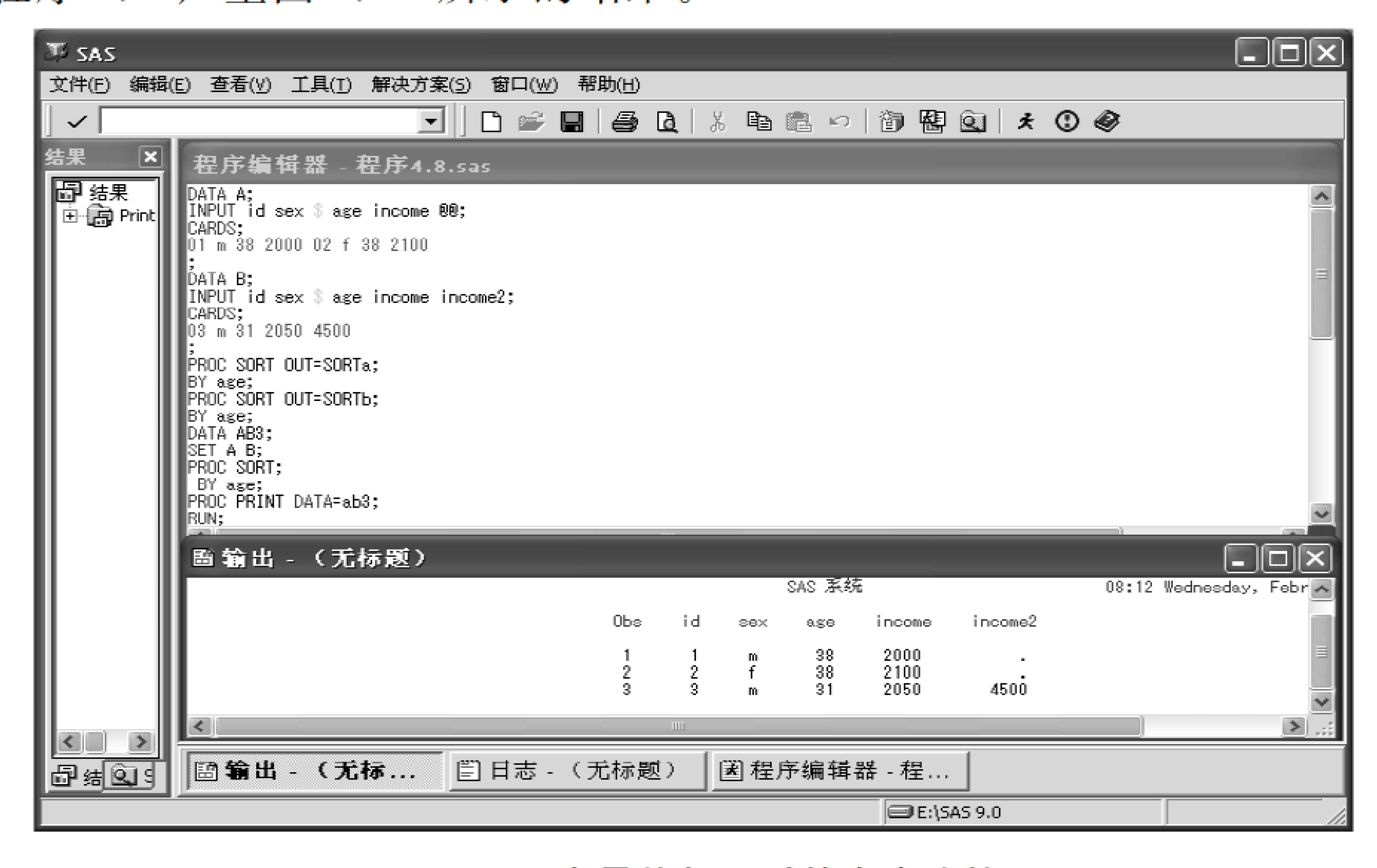


图 4.10 变量值相同时的个案连接

4.5 数据集"合二为一"

当每份问卷(个案)有几百个变量,而这几百个变量的数据在同一行输入不下时,则要考虑"一分为二"地拆为两半部分输入(当然也可分为两个记录输入),每半部分单独存储为一个数据集,那么,这个调查一共可建立两个数据集。但在统计分析时往往需要将这两个数据集"合二为一",才能对数据进行充分的利用。在这种情况下,要用 SAS 的 MERGE 语句(或用 SPSS 的 Data 菜单中的 MERGE FILE 命令)对数据集"合二为一"。这里所说的数据集"合二为一"其实可以同时对多达 50 个数据集横向合并变量。

MERGE 语句格式:

```
MERGE 数据集1  数据集2  … 数据集50;
```

4.5.1 按个案号配对合并变量

按个案号配对合并变量是指,a 数据集的 001# 个案,对应地与 b 数据集 001# 个案合并,而成为一个完整的 001# 个案。同理,a 数据集的 002# 个案,对应地与 b 数据集 002# 个案合并,而成为一个完整的 002# 个案,直到所有的个案都配对合并完毕。这种合并是变量的累加合并但观测值(个案)数目不变。

合并时,如果某个数据集 a 的个案数目少于另一个数据集 b 的个案数目时,短缺的个案上的所有变量则自动被赋予"."表示数据缺失值。

例 12:如表 4.1 所示,有数据集 a 和数据集 b,个案数目不一样多,变量数目也不一样多。如果把这两个数据集进行个案一一配对合并,则有图 4.11 的结果,语句见程序 4.9a。

表 4.1 数据集

数据集 a			
OBS(观测值)	Id (个案号)	Sex(性别)	Age (年龄)
1	01	1	38
2	02	2	30
3	03	1	24
数据集 b			
OBS(观测值)	Id (个案号)	Sex(性别)	vt
1	01	1	8
2	02	2	10

程序 4.9a:综合举例(个案不同而且变量又不一样的情形)。

```
DATA A1;
INPUT id sex age @@ ;
CARDS;
```

```
01 1 38 02 2 30 03 1 24
;
DATA B1;
INPUT id vt @@ ;
CARDS;
01 8 02 10
;
DATA AB5;
MERGE A1 B1;
PROC PRINT DATA= ab5;
TITLE '按个案号一一成对合并变量';
RUN;
```

运行程序 4.9a 产生图 4.11 所示的结果。

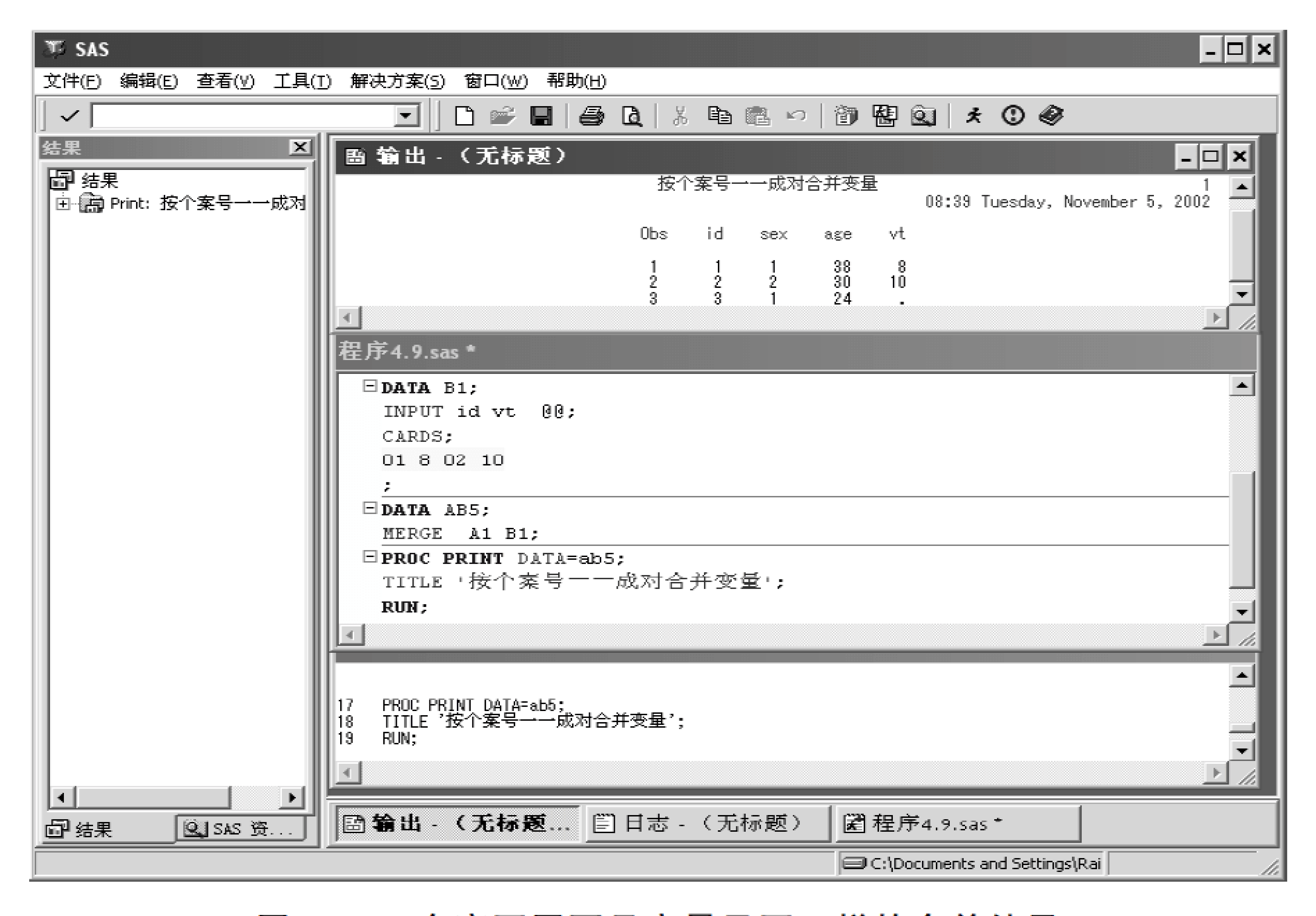


图 4.11 个案不同而且变量又不一样的合并结果

从图 4.11 看,合并时,如果某个数据集 B1 的个案数目少于另一个数据集 A1 的个案数目时,短缺的个案上的所有变量则自动被赋予“.”表示数据缺失值。

相反,尽管某变量在其他数据集里多次出现,则在新合并的数据集里它也只能出现一次,而且是最后出现的变量值覆盖前面出现的变量值。

例 13:尽管 sex 变量在其他数据集里也出现过,则在新合并后的数据集里它也只能出现一次,而且是最后出现的变量值覆盖前面出现的变量值。例如:

```
MERGE a b; /* 假定 a 和 b 两个数据集里都出现 sex 变量 */
```

那么,数据集 a 里的 sex 值,被数据集 b 里的那个 sex 值覆盖了。

程序 4.9b:数据集 A1 里的 sex 值,被数据集 B1 里的那个 sex 值覆盖了。

```
DATA A1;
INPUT id sex age @@ ;
CARDS;
01 1 38 02 2 30
;
DATA B1;
INPUT id sex score1 score2 @@ ;
CARDS;
01 1 78 88 02 1 86 95
;
DATA AB5;
MERGE A1 B1;
PROC PRINT DATA= AB5;
TITLE '按个案号一一成对合并变量';
RUN;
```

运行程序 4.9b 产生图 4.12 所示的结果。



图 4.12　个案相同但变量不一样时的合并

从图 4.12 看,数据集 A1 里的 sex=2 值,被数据集 B1 里的 sex=1 值覆盖了,从而只有 sex=1 的个案了(见图 4.12 所示的输出窗口)。

注意:一对一的个案配对合并,是按照个案号配对合并的。如果个案号不同,则不能采用 MERGE 语句进行配对合并,而必须改为“匹配合并”法。见下面的 4.5.2 节。

4.5.2 用 BY 语句进行匹配合并

上面所述是对个案合并或对变量合并，是简单的连接而已。但有时需要以某个变量为基准，将几个数据集的个案排列起来然后聚合为一个大的数据集。例如，以 location 变量为基准进行数据集的匹配合并。如表 4.2 和表 4.3 所示，有两个不同的数据集。要求按照地区升序排序后匹配合并成图 4.13 所示的数据集。

表 4.2 东城区数据集 a

OBS(观测值)	Id (个案号)	Sex(性别)	Age (年龄)	Location(地区)
1	01	1	38	Dc
2	02	2	30	Dc

表 4.3 东城区数据集 b

OBS(观测值)	Id (个案号)	Score1(数学成绩)	Score2(英语成绩)	Location(地区)
1	01	78	88	Hd
2	02	86	95	Cy
3	03	99	78	Xc



图 4.13 按照地区(location)升序排序后匹配合并

产生图 4.13 的语句见程序 4.10。

程序 4.10：两个数据集有一个公共变量 location。

```
/* 程序 4.10* /
DATA A1;
INPUT id location $  sex age ;
CARDS;
01 dc 1 38
02 dc 2 30
;
DATA B1;
INPUT id location $  score1 score2 ;
```

```
CARDS;
01 hd 78 88
02 cy 86 95
03 xc 99 78
;
PROC SORT DATA=a1;BY location; PROC PRINT;
;
PROC SORT DATA=b1;BY location; PROC PRINT;
;
DATA AB5;
SET A1 B1;
PROC SORT DATA=AB5;
BY location;
PROC PRINT;
TITLE '按 location 升序排序合并变量和个案';
RUN;
```

4.6 用 FILE 语句控制输出文件

1. 命令语句

```
FILE 文件说明 [选项];
```

2. 格式说明

FILE：它必须与 PUT、RETURN 语句连用。

文件说明：文件标签，用一对单撇号''括起来。

选项：控制输出行的当前行号，以及当前列位置及其最大长度。

如果不指定 FILE 语句，则由 PUT 语句将输出结果显示在日志(LOG)窗口。可用 FILE 语句的选项，指定变量应该在当前行号，以及当前列位置。并说明每个新输出页的开头想打印什么内容。

FILE 语句要与 PUT 语句连用，用以建立和控制某个输出行写到外部文件中。

3. 文件说明

路径和文件名：如 FILE 'C:\新 SAS\OUT1'。注意：路径和文件名要用一对单撇号''括起来。

LOG：日志窗口。如 FILE LOG。将 PUT 语句的输出显示在 SAS 的日志窗口中，默认 WORK. LOG。

PRINT：如 FILE PRINT N=PS; /* 把 PUT 语句的输出(包括 SAS 的所有输出)显示在输出窗口(OUTPUT) */

4. 选项说明

选项是控制输出行要存储在输出文件的什么位置。

(1) COLUMN＝变量名 a：SAS 自动把当前列的位置赋予变量 a。

(2) LINE＝变量 b：变量 b 的范围由 1～N 值。如果不指定 N＝值，则 LINE＝1。

(3) LS(或 LINESIZE)＝值：为报表指定变量在行上的长度，以及文件的最大长度。例如：

```
FILE PRINT LS=80;
```

如果用 PUT 语句指定的行长，大于由 LS＝80 所指定的行长，则每行将被截成两行或更多行。

例 14：

```
FILE PRINT LS=80;
PUT name $  1-40 location 41-81; /*  PUT 语句超过 80 行 */
```

那么，name 变量值将会输出在第 1 行，location 变量值将会输出在第 2 行。

(4) N＝PS(或 N＝PAGESIZE)或 N＝值：指针每次移动(显示)的行数。

例 15：

```
FILE PRINT LS=80 N=3;
…
PUT _PAGE_;
```

由于例 15 中 N＝3，所以指针由第 1 行移动到第 3 行，执行命令语句后返回到第 1 行，即 1～3 行有效。指针再移到第 4 行时，4～6 行有效。以此类推。

当指针移到 N＝PS 所指定的值，或移到“PUT _PAGE_;”语句时，指针便右移到下一页的首行。

如果当前输出文件是 PRINT 名，则必须是 N＝1，或 N＝PAGESIZE，见例 16。

例 16：产生一个两栏的通信录。每行含有性别 sex 和电话 dh 两栏内容，见程序 4.11。

程序 4.11：

```
DATA TX;
INPUT name $  10. location $  20. ;
CARDS;
Zhangsan Dongchengqu 86542334
Lisiguang Xichengqu
Wanglaowu Chaoyangqu
Malaoliu Haidianqu
;
DATA TXL;
FILE PRINT N=PS;        /* 宣告页中各栏的每行输出都能够用指针自由控制 */
  DO C=1,60;            /* 当 DO 循环到 END 语句时，指针移到第 60 列 */
    SET TX;             /* 复制并读取已有的数据集 TX */
    DO L=1 TO 48;       /* 当 DO 循环到 END 语句时，则每页输出两栏，每行 48 个个案的通信录 */
    PUT #1 @C name 10. +2 location 20.;   /* 指针移到首行首列位置输出 name 值，右移 2 列输出
```

```
                          location值 * /
    END;
  END;
PUT _PAGE_;           /* 执行"PUT _PAGE_;"语句,进入下一页。下一次执行 DATA语句时 L和 C的值
                         又都从新页的首行首列开始赋值 * /
PROC PRINT DATA= TXL;
```

运行程序 4.11 产生图 4.14 所示的结果。



图 4.14　产生一个两栏的通信录

程序 4.11 说明：

- N＝PS：宣告页中各栏的每行输出都能够用指针自由控制。
- SET 语句：复制并读取已有的数据集 TX。
- C＝和 L＝：指定指针的当前行和当前列位置。
- ＃L 和@C：让 PUT 语句在当前行和当前列位置输出 name 变量值及 location 变量值。行值由 1 递增到 48 行。
- 当内层 DO,循环到 END 语句时,指针移到第 60 列,输出下一页的 48 行。当外层 DO,循环到 END 语句时,则每页输出两栏,每行 48 个个案的通信录。此时执行"PUT _PAGE_;"语句,进入下一页。下一次执行 DATA 语句时 L 和 C 的值又都从新页的首行首列开始赋值。

4.7　OUTPUT 语句

OUTPUT 语句让 SAS 将当前个案的处理结果输出到指定的数据集里。

4.7.1 OUTPUT语句格式

```
OUTPUT 数据集1  数据集2;
```

数据集可以指定多个。如果不指定数据集名称,结果则被存入 DATA 语句所建立的数据集里。

简单的 DATA 步不必指定 OUTPUT 语句。因为 SAS 回到 DATA 步执行下一个 RUN 语句之前,会自动输出结果。

有以下控制形式时需要指定 OUTPUT 语句:

(1) 从每个数据行中建立两个或多个个案。

(2) 从一个输入的数据集里,创建两个或多个数据集。

(3) 将几个个案合并为一个个案时,在 DATA 步若有 OUTPUT 语句,则输出到所指定的数据集里而不显示在"输出(OUTPUT)"窗口。

4.7.2 一个个案的变量分几行输出

例 17:要求将每个人的三门功课分三行输出,每行只输出一门的成绩,命令语句见程序 4.12。

程序 4.12:

```
DATA outp;
INPUT id score1-score3 @@ ;
 DROP score1-score3;
     score=score1;
     OUTPUT;
     score=score2;
     OUTPUT;
     score=score3;
     OUTPUT;
CARDS;
01 88 90 92 02 78 88 68
;
PROC PRINT;
TITLE '个案一分为三';
```

运行程序 4.12 产生图 4.15 所示的结果。

4.7.3 一个 DATA 步创建多个数据集

例 18:要求将每个人的数据存储为两个数据集。如 25 岁以下的人存储成一个数据集 A,25 岁及 25 岁以上的人存储成另一个数据集 B,命令语句见程序 4.13。

程序 4.13:

```
DATA A B;
```

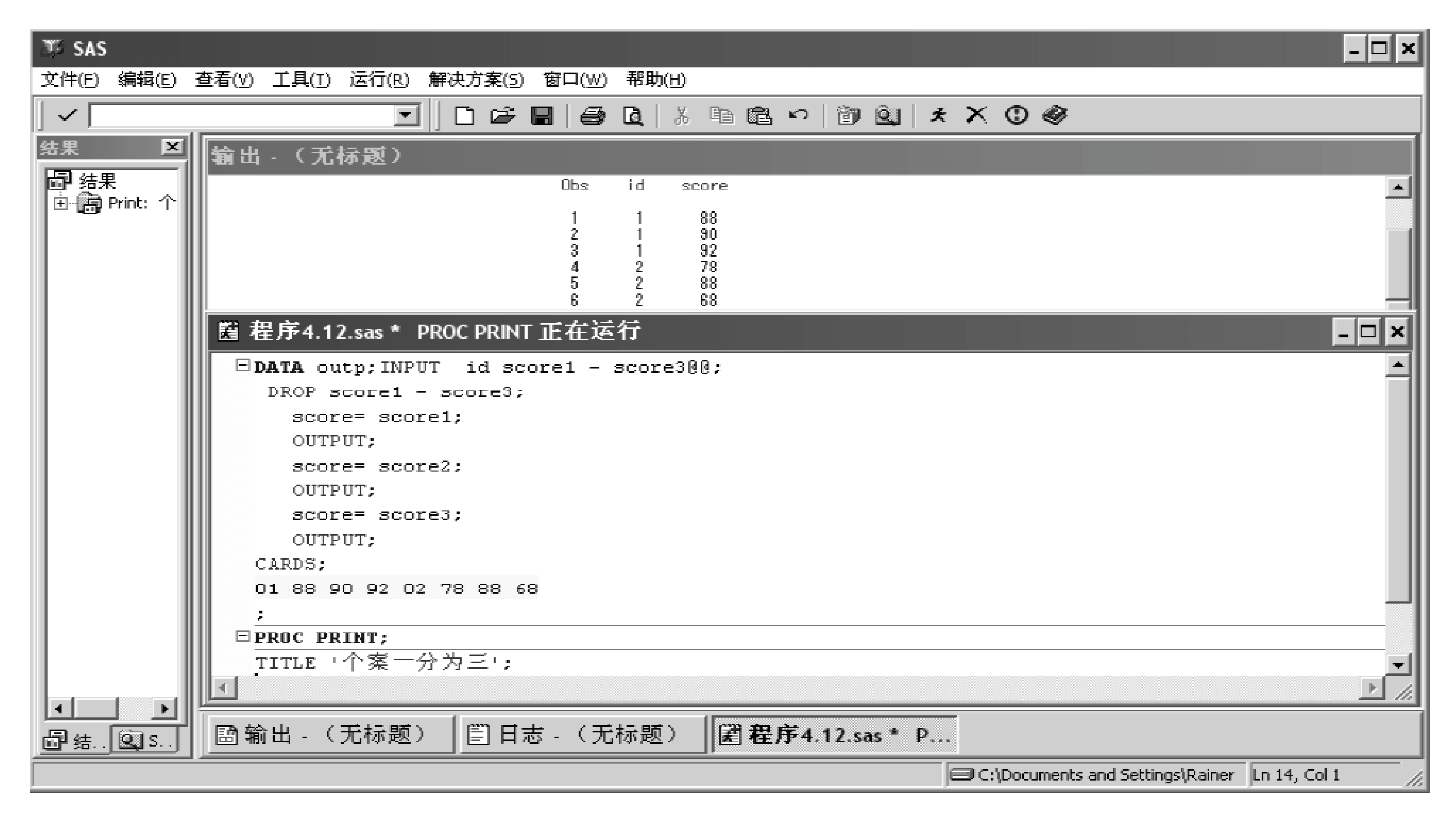


图 4.15 每行变成只输出一门的成绩

```
INPUT id sex $  age h w @ @ ;
IF age< = 25 THEN OUTPUT A;        /* 将 25 岁以下的人存储成一个数据集 A* /
PUT DATA= a;
ELSE OUTPUT B;                     /* 将 25 岁及以上的人存储成一个数据集 B* /
CARDS;
01 m 20 1.68 65.5 02 f 25 1.70 68.5
03 m 28 1.71 69.5 04 f 30 1.69 59.5
;
PROC PRINT DATA= A;
TITLE;
PROC PRINT DATA= B;
RUN;
```

运行程序 4.13 产生图 4.16 所示的结果。

4.7.4 多行信息合并为一行

例 19：将每个人几次的收支累计起来输出为一行，见程序 4.14。

程序 4.14：

```
DATA UnTotal;
INPUT id pay1;
CARDS;
01 105.50
01 78.30
02 110.50
02 35.50
03 115.00
```

```
;
PROC SORT DATA= UnTotal;
 BY id;
PROC PRINT;
TITLE '支出未累计';
DATA Total;
SET UnTotal;
 BY id;
IF FIRST.id THEN pay2= 0;
  pay2+ pay1;
  DROP pay1;
IF LAST.id THEN OUTPUT;
 PROC PRINT;
TITLE '支出累计';
```

运行程序 4.14 产生图 4.17 所示的结果。



图 4.16 将原始数据存储为两个数据集(见“输出”窗口)

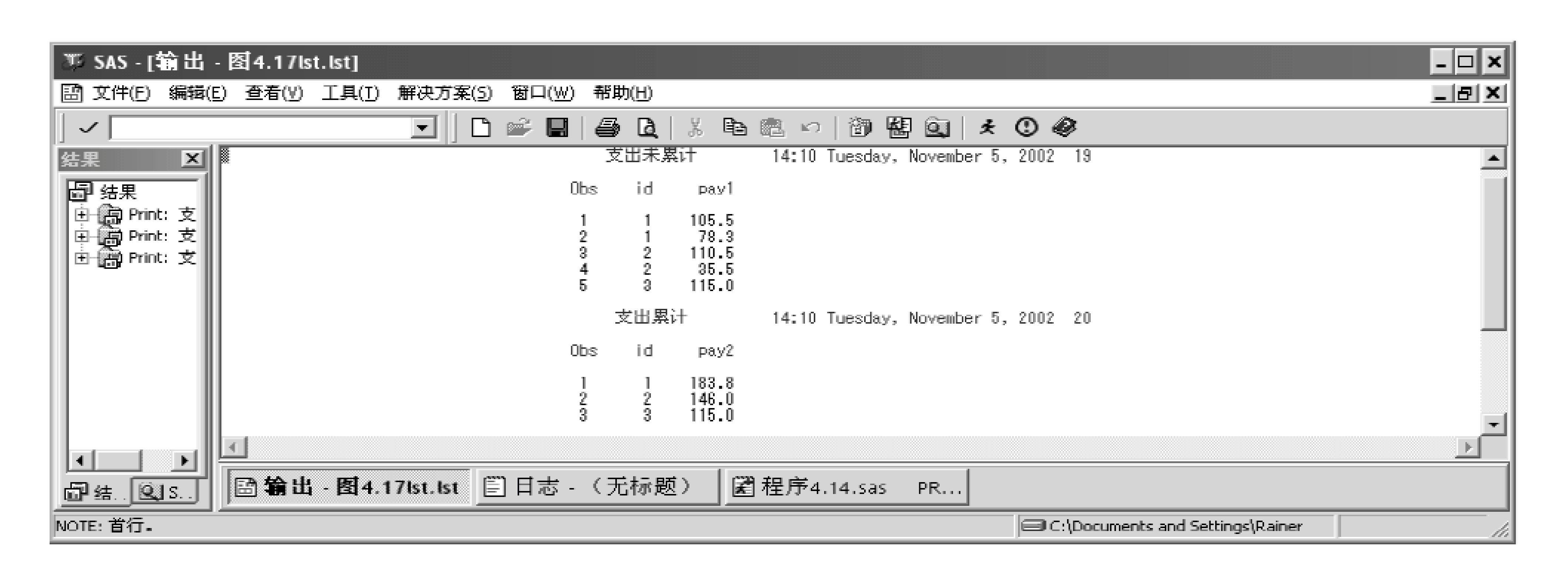


图 4.17 将每个人几次的收支累计起来输出为 1 行

从图 4.17 中的“支出未累计”与“支出累计”两个数据集的比较可以看出，已由原来的 5 行观测值累计合并为 3 行、每行显示 1 个人的总支出。

4.8 用 DATASETS 过程修改数据集

DATASETS 过程可对数据库文件进行追加、复制、列表、改名或删除。还可改变 DATA 步的变量名、变量类型、变量长度或标记。

1. DATASETS 过程命令

```
PROC DATASETS [LIBRARY=库逻辑名]; /* 若不指定库逻辑名,则默认为“WORK.*”*/
  APPEND BASE=前面的数据集 DATA=后随的数据集;
  MODIFY 数据集;          /* 一个 DATASETS 过程允许指定多个 MODIFY。MODIFY 语句后面可以指定
                             FORMAT、INFORMAT、LABEL、RENAME 等子命令。数据集只能指定 1个 */
  FORMAT 老变量 新格式;   /* 改变或取消由 MODIFY 所指定的数据集里的变量属性。如果不指定
                             新格式,则取消原格式 */
  INFORMAT 老变量 新格式;/* 改变由 FORMAT 所指定的变量输入的属性 */
  LABEL 变量=新标签;      /* 改变由 MODIFY 所指定的变量标签 */
  RENAME 老变量名 新名;
```

2. 子命令说明

APPEND BASE＝前面的数据集 DATA＝后随的数据集：如果默认尾随的数据集，则追加当前工作区中的数据集。

例 20：用 APPEND 子命令在数据集 a 后面追加数据集 b，见程序 4.15。

程序 4.15：

```
DATA A;
INPUT id sex $  age income @@ ;
CARDS;
01 m 38 2000 02 f 30 2100
;
DATA B;
INPUT id sex $  age income ;
CARDS;
03 m 31 2050
;
PROC DATASETS;                /* 主过程名 */
    APPEND BASE=a DATA=b;     /* 在数据集 a后面追加数据集 b*/
 MODIFY b;                    /* 调用工作区里的数据集 WORK.b进行修改 */
 INFORMAT income 6.0;         /* 把变量 income修改为 6.0的长度 */
PROC PRINT;
RUN;
```

运行程序 4.15 产生图 4.18 所示的结果。



图 4.18　用 APPEND 子命令在数据集 a 后面追加数据集 b(见“日志”)窗口

4.9　查阅数据集的信息

如果想查阅数据集里的变量、变量类型、变量值等属性,可用 PROC CONTENTS 过程命令。

1. 命令格式

```
PROC CONTENTS DATA=数据集 a;    /* 查阅数据集 a 里的信息 */
PROC CONTENTS;                  /* 查阅当前工作区里的数据集信息 */
```

2. 例子

查阅当前工作区里的数据集信息,见程序 4.16。

程序 4.16:

```
DATA A;
INPUT id sex $  age income @@ ;
CARDS;
01 m 38 2000 02 f 30 2100
;
DATA B;
INPUT id sex $  age income ;
CARDS;
```

```
03 m 31 2050
;
PROC DATASETS;
   APPEND BASE=a DATA=b;
   MODIFY b;
   INFORMAT income 6.0;
PROC CONTENTS ;
```

运行程序 4.16 产生图 4.19 所示的结果。



图 4.19 在输出窗口查阅当前工作区里的数据集信息

习 题 4

1. 为什么要用 LIBNAME 语句指定路径？试举例。
2. 什么是分组控制？
3. 试举一个分组控制的例子。
4. 试举一个用 MERGE 语句将数据集“合二为一”的例子。
5. 试举一个“将每个人几次的收支累计为 1 次”的例子。

第 5 章 chapter 5

数据的定义及汉化

数据挖掘与统计分析之前,需要在程序中使用一些基本的过程命令。本章介绍程序中最基本的 SAS 过程命令及语句,是为以后的数据挖掘奠定基础。

5.1 DATA 语句

DATA 语句格式:

```
DATA 数据集;                   /* 例如:用"DATA score;"语句定义数据集 work.score */
或 DATA a b;                    /* 定义数据集 work.a 和数据集 work.b */
或 LIBNAME LB2 'F:\新 SAS';     /* 定义 F:\新 SAS 路径 */
DATA LB2.test2;                 /* 建立永久的数据集 F:\新 SAS\test2.SAS7BDAT */
或 DATA _NULL_;                 /* 为了提速,可以不建立数据集 */
```

5.2 INFILE 语句

1. 语句格式

```
INFILE 'D:\新 SAS\my.txt'; /* 从 D:\新 SAS\子目录或文件夹读取 my.txt 纯 ASCII 码数据 */
```

2. 说明

实际应用时通常会有成百上千个个案。这么多的数据行夹杂在命令语句里显得不便,也容易丢失。因此人们往往将成百上千个个案的数据单独编辑存储为一个纯文本的数据文件(如 my. txt)。统计分析时用 INFILE 'D:\新 SAS\my. txt'命令把数据文件调到内存中即可。

纯文本的数据文件是 ASCII 码数据,在任何计算机系统中都可以阅读,并且便于修改。纯文本的数据一般是在 Windows"附件"的"记事本"中编辑和存储,起扩展名为 *. txt。如 D:\新 SAS\子目录中有一个 my. txt 纯文本的数据文件,其数据行如下:

```
001 m 30 3000
```

```
002 f 28 2800
003 m 35 3200
004 f 32 3100
```

下面就可以用 INFILE 语句将 my.txt 中的数据读取到 SAS 工作区中进行处理，见程序 5.1。

程序 5.1：

```
LIBNAME LB 'D:\新 SAS';
DATA LB.income;
INFILE 'D:\新 SAS\my.txt'; /* 从 'D:\新 SAS\'路径中读取 my.txt 纯文本的数据 * /
INPUT id sex $  age height;
PROC PRINT;
RUN;
```

运行程序 5.1 产生图 5.1 所示的结果。



图 5.1 输出结果

3. 注意事项

在 my.txt 数据文件里，每一行的数据有 80 列。但在“INPUT id 1-5 sex $ 10 … height 77—78;”语句中最后一个变量的栏目位置分配到 78 列即止。此时读到数据文件的第 78 列后有可能不再读取，这时便丢失了最后两列的数据。或者有可能把最后两列数据错判给下一个个案。这就会引起读取数据的错位和错行。即读取数据一错百错。因此，定义固定栏目格式的变量值时，一定要与数据文件中的变量值的真实位置一一对应，绝对不能错位。

5.3 INPUT 语句

INPUT 语句用以定义数据集里的变量、变量类型、变量值、小数位等属性。它相当于 SPSS 系统中的 DATA LIST 命令。

5.3.1 用 INPUT 语句定义固定格式的变量

1. 语句格式

```
INPUT  v1  栏目位置  v2  栏目位置  v3  栏目位置;
```

例如：

程序 5.2：

```
DATA A;
INPUT id 1-2 sex $  4 age 6-7 income 9-13       .1;
CARDS;
01 m 38 20000 /* 数据只能分行写,不能像自由格式那样在一行上写几个个案 */
02 f 30 21000
;
PROC PRINT;
```

2. 固定格式的写法

例如："id 1-2"的写法，先写变量名如 id，空出一列后，写起始栏位如"1"，加一个"-"英文减号，再写终止符如"2"。

"sex $ 4"表示：sex 变量是字符型的，其值应该输入在每行的第 4 列上。

"income 9-13   .1"表示：收入变量的值，应该输入在每行的第 9～13 列上，而且小数位占 1 位。

5.3.2 用 INPUT 语句定义自由格式的变量

1. 语句格式

```
INPUT v1 v2 v3 @@ ;
```

例如：

程序 5.3：

```
DATA A;
INPUT id sex $  age income 5.1 @@ ; /* 自由格式可用@@ 连读几个个案 */
CARDS;
01 m 38 20000 02 f 30 21000 /* 自由格式允许在一行上写几个个案 */
;
PROC PRINT;
```

2. 自由格式的写法

例如"id"的写法，只写变量名如 id，栏位统统省去。

"sex $"表示：sex 变量是字符型的。

"income 5.1"表示：收入变量的值的长度为 5 列，小数位占 1 位。

@@：自由格式可用@@连读几个个案。

5.3.3 用INPUT语句指定格式化的输入方式

此格式，在变量名的后面不写栏位，但要写变量值的长度以及小数位。

1. 语句格式

INPUT 变量1 长度.小数 变量2 长度.小数 …；

2. 说明

变量1 长度.小数：如“INPUT age 3.1；”表示年龄有3位长度，其中小数位占1位，见程序5.4。

程序5.4：

```
DATA abc;
INPUT id 3. sex $  age 4.1 ;
CARDS;
001 m 405 002 f 410 003 m 505
;
PROC PRINT ;
```

或

```
DATA abc;
INPUT id 3. sex $  age ;
CARDS;
001 m 40.5
002 f 41.0
003 m 50.5
;
PROC PRINT ;
```

运行程序5.4产生图5.2所示的结果。



图5.2 自由格式的输出

5.3.4　INPUT 语句含有挖掘功能

这里所说的挖掘，既可以把指针指向第几行去挖掘数据，又可跳到第几列去读取数据，还可将指针空移几列（如＋3）后再去读取数据。

1．语句格式

```
INPUT @ 3 id1 + 1 sex + 2 age @ c location $  8. # 2 id2 3. ;
```

2．说明

@3：绝对指针跳到第 3 列去读取个案号 id1。

＃2：指针跳到第 2 行去读取数据。

@c：指针跳到有字符 c 标志的变量上去读取数据。

id2 3.：id2 变量值的长度为 3 列字符，没有小数位。

程序 5.5：按自己的风格输出数据集的内容。

```
DATA as;
  S= 10;
INPUT id1 @ 4 sex $  age @ s location $  11. # 2 id2 3. + 1 income 6.;
CARDS;
01 m 40 Dongchengqu
001 04800
02 f 42 Xichengqu
002 05000
03 m 50 Haidianqu
003 10000
;
PROC PRINT;
```

运行程序 5.5 产生图 5.3 所示的结果。

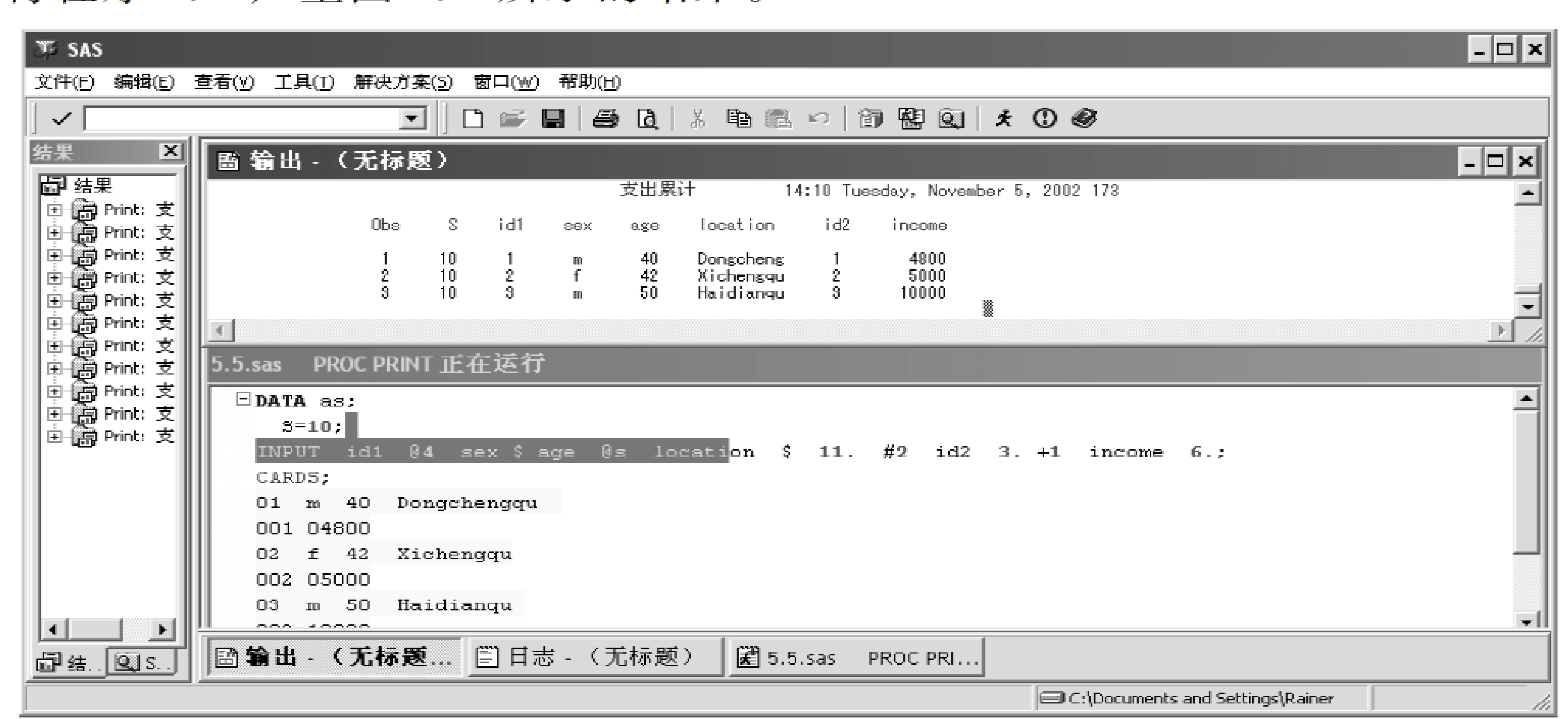


图 5.3　控制指针挖掘的技巧

5.4 用 LABEL 语句定义变量标签

在 INPUT 语句中,定义的变量名的长度不能大于 8 个字符。因此,对于大于 8 个字符的变量名缩写为 a、b、c 或 v1、v2、v3 之类的变量名,之后可用 LABEL 语句将此简名标识为原名或者用汉字汉化更美观。

1. LABEL 语句格式

LABEL 变量 1= '标签' 变量 2= '标签' ……;

2. 例子

程序 5.6:

```
DATA;
INPUT id sex age@@ ;
LABEL sex= '性别' age= '年龄' id= '个案号';
CARDS;
01 1 26 02 2 30 03 1 40
;
PROC PLOT;
  PLOT sex * age;
PROC PRINT;
RUN;
```

运行程序 5.6 产生图 5.4 所示的结果。

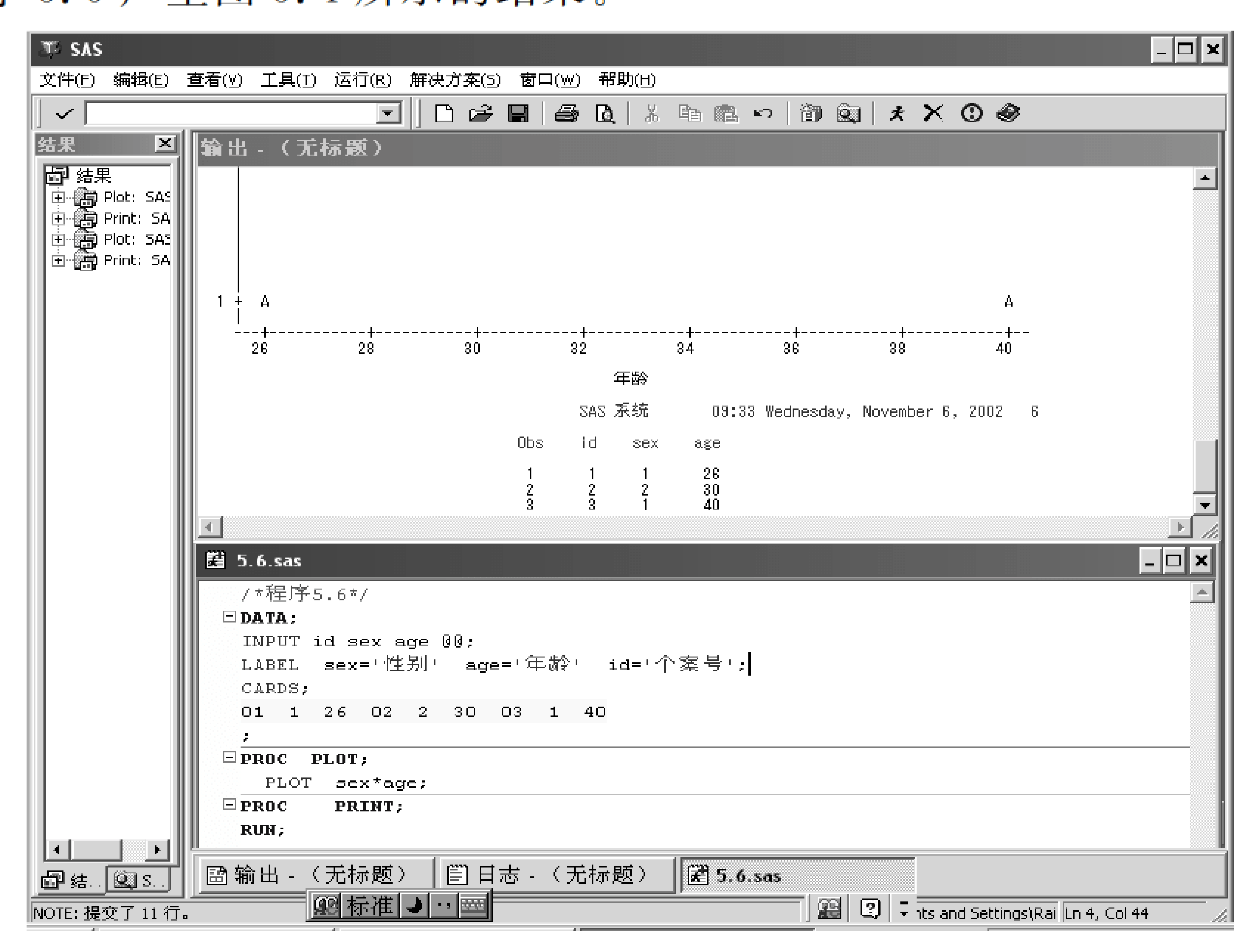


图 5.4 用 LABEL 语句汉化了变量名称

5.5 用 FORMAT 及 VALUE 语句定义数值标签

FORMAT 语句可用于指定变量值标签，即数值标签。FORMAT 语句的第 2 个功能是指定“格式化输入”，即指定某个变量值的输出长度。

5.5.1 定义数值标签

要用 FORMAT 与 VALUE 语句共同定义数值标签。

1. 语句格式

```
PROC FORMAT;
  VALUE aF 1='标签' 2='标签' bF 1='标签' 2='标签' 3='标签';
FORMAT a=aF.;   /* aF.是新建的变量,a是在 DATA步的旧变量 */
FORMAT b=bF.;   /* 将 bF.新变量的值,赋予旧变量 b*/
```

2. 例子

程序 5.7：

```
DATA;
INPUT id sex age @@ ;
LABEL sex='性别' age='年龄' id='个案号';
CARDS;
01 1 26 02 2 30 03 1 40
;
PROC FORMAT;
  VALUE sexF 1='男性' 2='女性' ;
FORMAT sex sexF.;
PROC FREQ;
  TABLE age * sex;
PROC PRINT;
RUN;
```

程序 5.8：将年龄分为“老　中　青”3 个年龄段。

```
DATA v;
INPUT id sex age income @@ ;
CARDS;
01 1 30 3400 02 2 35 3500 03 1 50 4000 04 2 22 1800
;
PROC FORMAT;
    VALUE sexF 1='男性' 2='女性' ;
    VALUE ageF LOW-40='青年人' 40-60='中年人' 60-HIGH='老年人';
    VALUE incomeF LOW-2000=1 2001-3000=2 3001-HIGH=3;
```

```
FORMAT sex sexF. age ageF. income incomeF.; /* 使新格式生效 * /
LABEL sex='性别' age='年龄' id='个案号';
PROC FREQ;
  TABLE sex * age;
PROC PRINT;
RUN;
```

请上机运行程序 5.7 和程序 5.8 并分析输出结果。

5.5.2 指定“格式化输入”

遇到日期函数与日期时间的数据时，往往会输出从 1960 年 1 月 1 日以来的天数(或零点以后的秒数)。因此，必须用 FORMAT 语句对日期变量赋予相应的年月日或时间格式。

程序 5.9：

```
DATA d;
INPUT id sex DATE DATE7.;
FORMAT DATE DDMMYY8.;
CARDS;
01 1 28oct80
02 1 26mar90
03 2 01nov88
04 2 16jan84
;
PROC PRINT;
RUN;
```

运行程序 5.9 产生图 5.5 所示的结果。



图 5.5 DATE 栏下的日期时间变成了 28/10/80

从图 5.5 的输出窗口看,DATE 栏下的日期时间变成了 28/10/80。

5.5.3 用 FORMAT 语句指定变量值的格式

FORMAT 语句除了上述的两种格式外,还可用它来指定“变量名长度.小数位”。

1. 语句格式

FORMAT v1 格式 v2 格式;

2. 说明

(1) 如果某变量(如 sex)在几个 FORMAT 语句中出现,则它只能以在最后那个 FORMAT 语句中的格式为准。例如,在 DATA 步中用 FORMAT 语句给 sex 变量定义了格式“6.”,但在 PROC 步又用 FORMAT 语句给 sex 重新定义格式“5.”,那么,以最后这个 FORMAT 语句所指定的格式为准。

(2) 如果取消 DATA 步中已指定的输出格式,则在 FORMAT 语句中只需指定变量名,但不要指定格式。

3. 例子

程序 5.10:

```
DATA d;
INPUT id sex DATE DATE7.;
FORMAT DATE DDMMYY8.;
CARDS;
01 1 28oct80
02 1 26mar90
03 2 01nov88
04 2 16jan84
;
PROC PRINT DATA=D;
FORMAT DATE YYMMDD8.;
RUN;
```

运行程序 5.10 产生图 5.6 所示的结果。

比较图 5.6 和图 5.5 可以看出,时间格式后来被程序 5.10 中的“FORMAT DATE YYMMDD8.;”语句改为“年-月-日”(如 80-10-28)的格式。



图 5.6 DATE 栏下的日期时间变成了 80-10-28

5.6 用 TITLE 语句显示标题

1. 语句格式

TITLEn '标题的内容'; /* n=1～10。默认为“TITLE '标题的内容';” */

如果规定的标题比指定的行长还要长，标题则被分为几行输出。一旦规定了标题就一直生效，直到指定新的标题为止。

TITLE 语句应写在“PROC 过程名;”语句之后，下一个“PROC 过程名;”语句之前，或写在下一个 DATA(或 RUN)语句之前。

2. 例子

程序 5.11：

```
DATA A1;
INPUT id sex age @@ ;
CARDS;
01 1 38 02 2 30
;
TITLE '未成对合并变量'; /* TITLE 定义在两个 DATA 步之间 */
DATA B1;
INPUT id score1 score2 @@ ;
CARDS;
01 78 88 02 86 95
```

```
;
DATA AB5;
MERGE A1 B1;
PROC PRINT DATA= ab5;
TITLE '按个案号一一成对合并变量'; /* TITLE 定义在两个 PROC 步之间 * /
PROC MEANS;
RUN;
```

运行程序 5.11 产生图 5.7 所示的结果。



图 5.7　TITLE 语句应该写在两个过程名之间，或写在两个 DATA 步之间

程序 5.12：

```
DATA A1;
INPUT id sex age @ @ ;
CARDS;
01 1 38 02 2 30
;
PROC PRINT;
RUN;
TITLE '显示数据集 a1'; /* TITLE 定义在 RUN 之后，不是全程的标题而仅仅显示 RUN 后面过程的标
                        题 * /
PROC FREQ;
```

程序 5.12 请读者自行上机体验。

程序 5.13：

```
DATA A1;
```

```
INPUT id sex age @@ ;
CARDS;
01 1 38 02 2 30
;
PROC MEANS;VAR age;
PROC FREQ;
TITLE '显示输出结果'; /* TITLE 定义在两个 PROC 步之间 */
PROC PRINT;
```

运行程序 5.13 产生图 5.8 所示的结果。

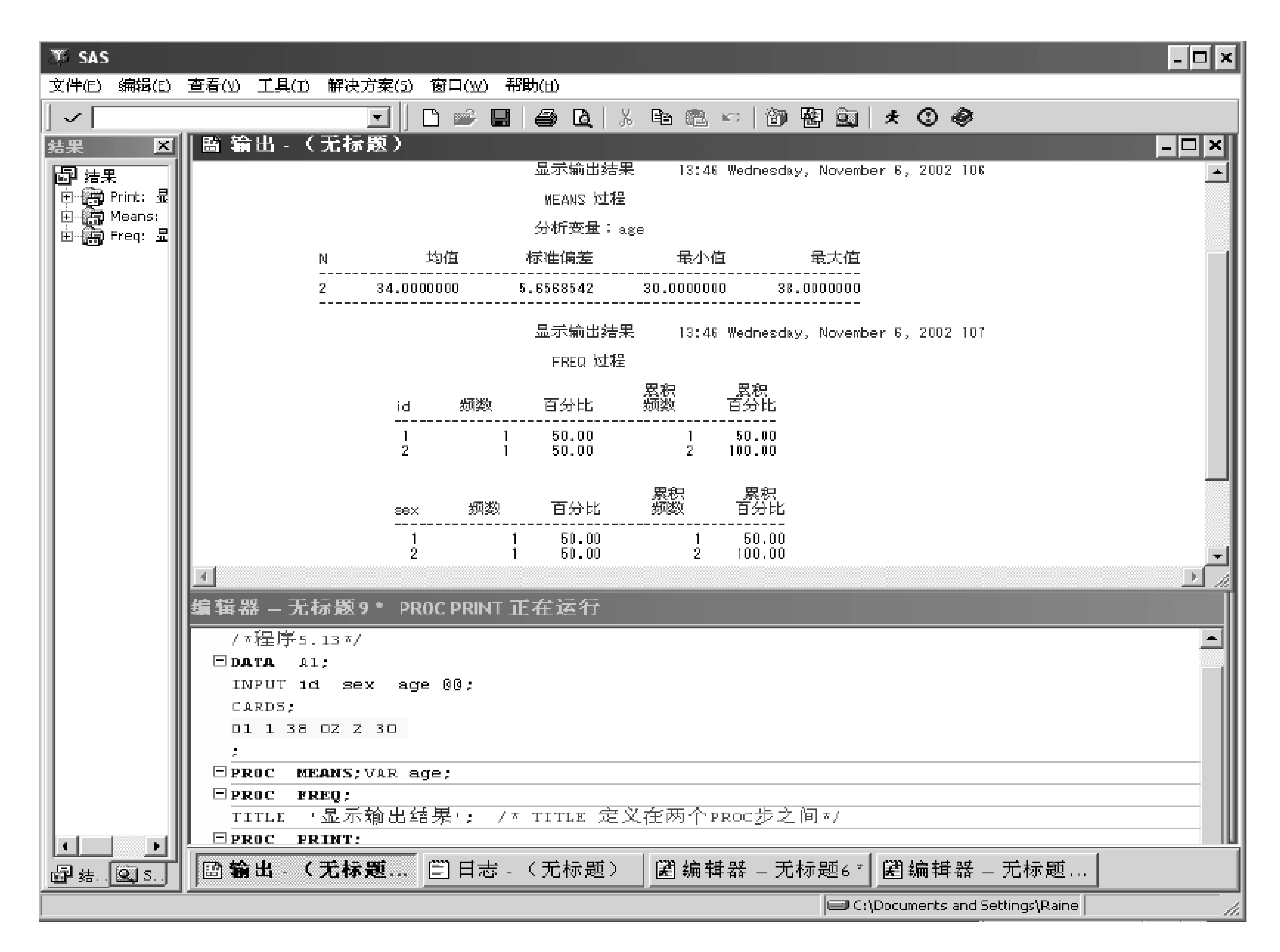


图 5.8 TITLE 定义在两个 PROC 步之间

从图 5.8 看,由于 TITLE 定义在两个 PROC 步之间,所以成了两个过程的公共标题。

5.7 数据挖掘常用的统计过程

在 SAS 系统中,统计过程是由专有名词 PROC 加以定义和执行。PROC 是 Procedure 一词的简写,即统计过程(命令)。

PROC 的命令格式:

```
PROC 统计过程名 [选项]; /* 选项有“DATA=数据集名”等 */
```

例如:

```
PROC MEANS DATA=a;
```

SAS 系统中,常用的统计过程有 FREQ、MEANS、UNIVARIATE、FACTOR、CLUSTER 等统计过程(命令)。下面将简要地介绍常用的统计过程(命令)的功能,让大

家有个感性认识。关于统计命令的详细功能及其数据分析，将在各章中详细介绍。

5.7.1 用 PROC FREQ 过程做简单的频数分布

```
PROC FREQ DATA=a;
TABLE 变量名; /* 例子见程序 5.14* /
```

程序 5.14：

```
TITLE ' ';
DATA A1;
INPUT id sex age @@ ;
CARDS;
01 1 38 02 2 30
;
TITLE '一般的频数统计';
PROC FREQ;
  TABLE sex;
RUN;
```

5.7.2 用 PROC CHART 过程画图

```
PROC CHART DATA=a;
  HBAR v1;
  VBAR v1; /* 例子见程序 5.15* /
```

程序 5.15：

```
TITLE ' ';
DATA A1;
INPUT id sex age @@ ;
CARDS;
01 1 38 02 2 30
;
TITLE '画出水平条形图';
PROC CHART DATA=a1;
  HBAR sex;
TITLE '画出垂直条形图';
  VBAR sex;
RUN;
```

5.7.3 用 PROC PLOT 过程画散点图

```
PROC PLOT DATA=a1;
  PLOT v1 * v2; /* 例子见程序 5.16a* /
```

程序 5.16a：画散点图。

```
DATA A1;
INPUT id sex age @@ ;
CARDS;
01 1 38 02 2 30
;
TITLE '画散点图';
 PROC PLOT DATA=a1;
  PLOT sex* age;
RUN;
```

程序 5.16b：画椭圆形等高线图。

```
DATA P2;
DO x=-3 TO 3 BY .1;
  DO y=-2 TO 2 BY .1;
   Z=SQRT(x*x+y*y);
   OUTPUT ;
  END;
 END;
PROC PLOT;
    PLOT y*x=z /CONTOUR HAXIS=-3 TO 3 BY .1
VAXIS=-2 TO 2 BY .1;
```

运行程序 5.16b 产生椭圆形等高线图，见图 5.9 所示的结果。

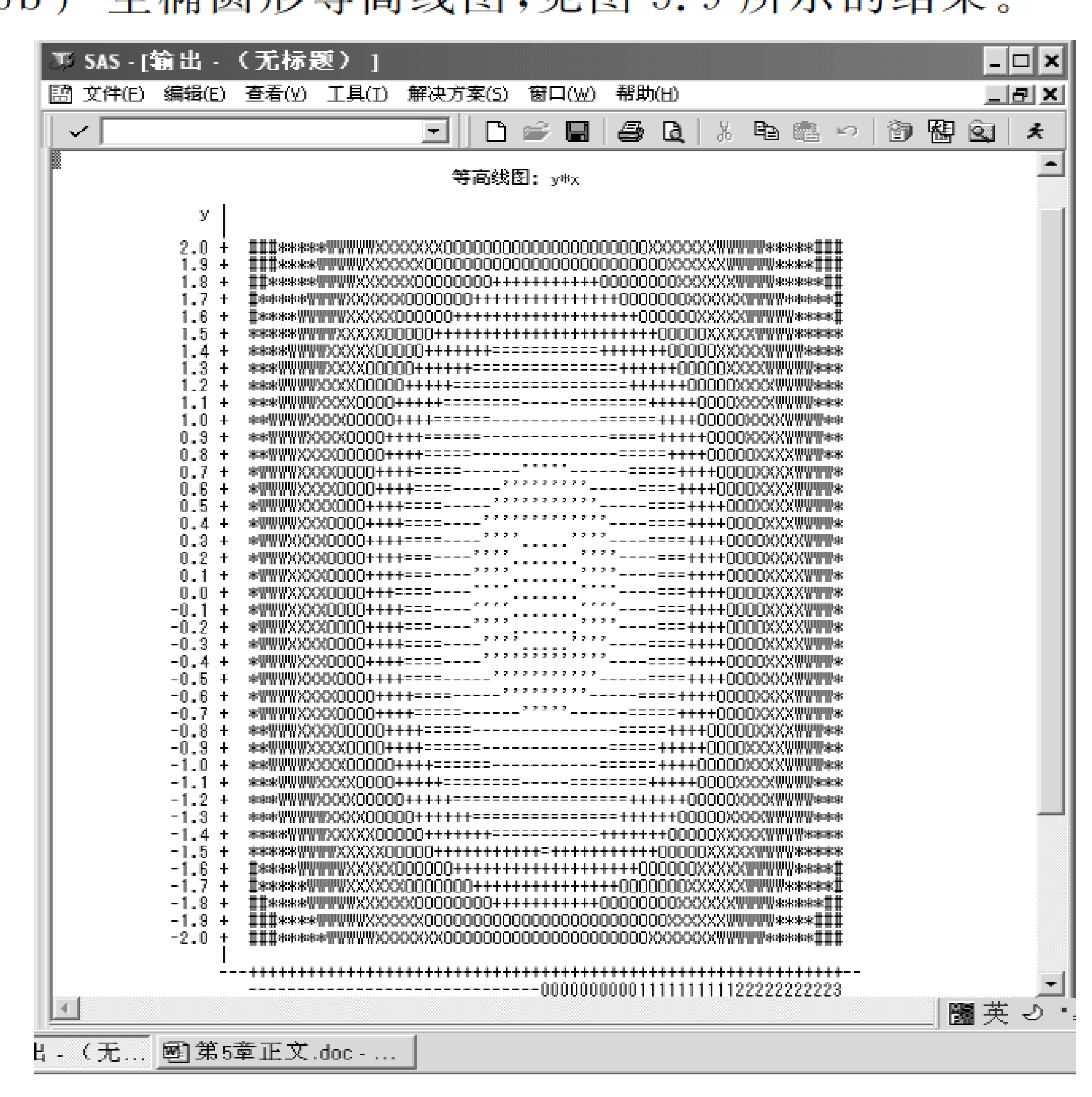


图 5.9 椭圆形等高线图

5.7.4 用 PROC MEANS 过程统计均值分布

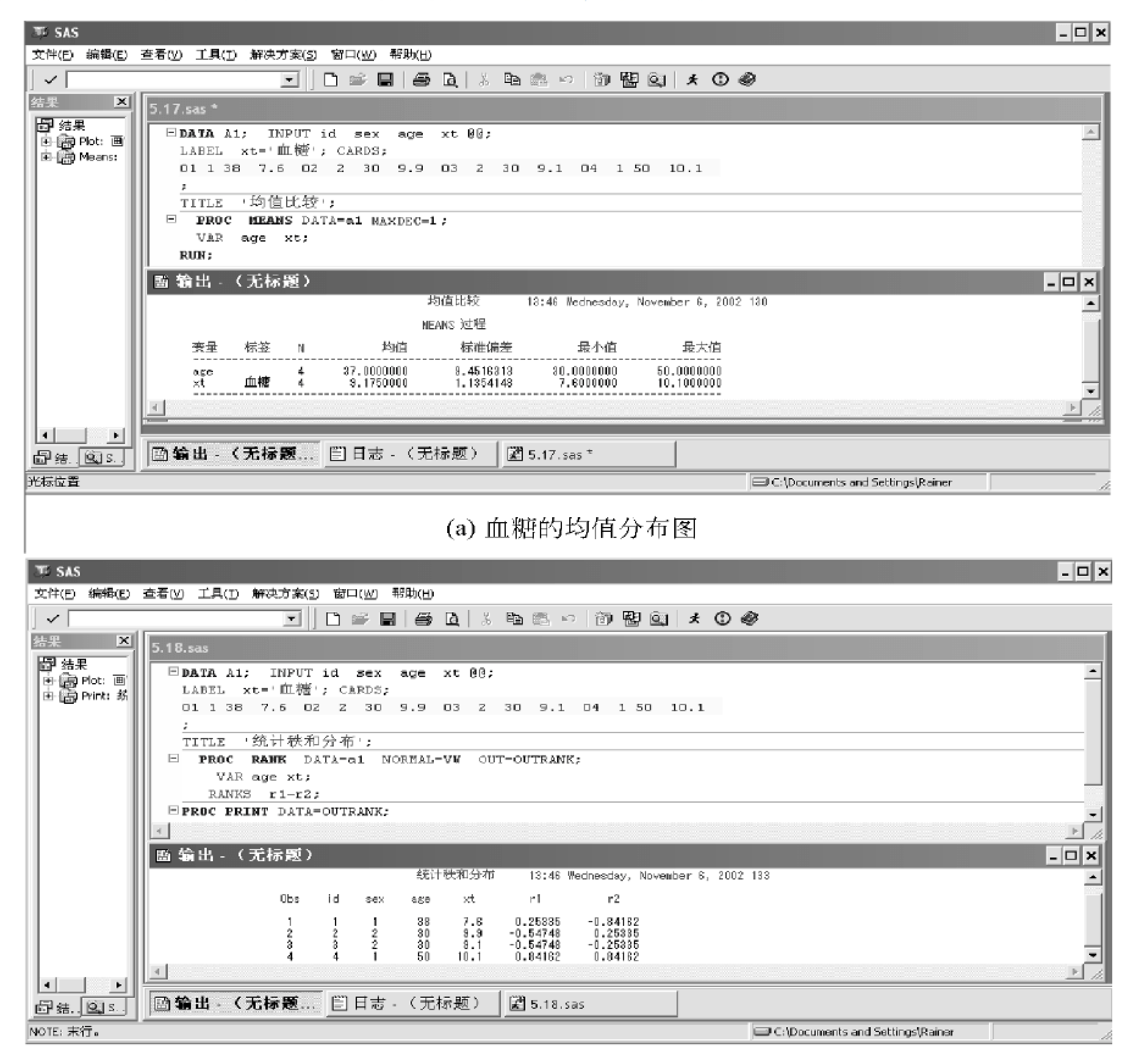
格式：

PROC MEANS DATA= a1 MAXDEC= 2; /* 例子见程序 5.17* /

程序 5.17：

```
DATA A1;
INPUT id sex age xt @@ ;
LABEL xt= '血糖';
CARDS;
01 1 38 7.6 02 2 30 9.9 03 2 30 9.1 04 1 50 10.1
;
TITLE '均值比较';
   PROC MEANS DATA= a1 MAXDEC= 1;
   VAR age xt;
RUN;
```

运行程序 5.17 产生图 5.10(a)所示的均值分布图。



(a) 血糖的均值分布图

(b) 输出秩分

图 5.10 分布图

5.7.5 用 PROC RANK过程统计秩和分布

格式：

```
PROC RANK DATA = a1 TIES= MEAN| HIGH|LOW NORMAL= VW
          OUT= OUTRANK;   /* VW 即 Van Der Waerden * /
          VAR v1 v2;
   RANKS 秩名(如 RANKsex RANKeduc);
   BY 变量(如 location);   /* 按 location分组进行秩分变换前必须先排序 * /
```

说明：当无法确定数据的总体分布，或当数据呈现明显的偏态时，或数据仅仅是顺序尺度(仅以程度表示无具体数值)时，都不宜采用参数统计，而必须采用不依赖于某种总体分布的统计法，即只能比较其分布，而不能比较参数。这是非参数统计。

这时应将原始数据进行“秩得分”变换：先将变量值从小到大(或从大到小)进行排序，然后分配序号，使成为原始数据的秩次。

例子见程序 5.18。

程序 5.18：

```
DATA A1;
INPUT id sex age xt @@ ;
LABEL xt= '血糖';
CARDS;
01 1 38 7.6 02 2 30 9.9 03 2 30 9.1 04 1 50 10.1
;
TITLE '统计秩和分布';
  PROC RANK DATA= a1 NORMAL= VW OUT= OUTRANK;
    VAR age xt;
    RANKS r1-r2;
PROC PRINT DATA= OUTRANK;
TITLE ' NORMAL= VW,VW正态法';
PROC PLOT DATA= a1;
  PLOT age * r1 xt * r2 ;
RUN;
```

运行程序 5.18 后产生的部分输出见图 5.10(b)。

5.7.6 用 PROC TABULATE 制表

格式：

```
PROC TABULATE DATA= a1 FORMAT= 14.2(或其他值);
    CLASS 定类变量;    /* 必须是定类型变量 * /
    VAR v1 v2;         /* 必须是定距定比型变量 * /
     FREQ 变量 v;      /* 变量可省略 * /
TABLE sex ALL,age(N MAX MIN MEAN);   /* sex 在行上, age 在列上 * /
```

或

```
TABLE 页变量, sex ALL,age(N* F=6. PCT* F=10.2 MAX MIN MEAN);
        /* 页变量决定了分页。sex 在行上, age 在列上。N* F=6.表示个案数占 6 位,其中小数
           位占 0 位 */
        /RTS=15 BOX='平均成绩'; /* RTS=15是行标题长度。BOX 显示横行的标题 */
```

例子见程序 5.19。

程序 5.19:画出学生成绩表。有 3 班学生,每人的成绩有 3 门(score1-score3)。要求计算每人的平均成绩(变量 Average),其中 90 分以上为 A 类,71—89 分为 B 类,70 分以下为 C 类。

```
DATA score;
OPTION PS=48;
INPUT id $  score1-score3 @@ ;
classid=SUBSTR(id,1,3);
Average=MEAN(OF score1-score3);
IF Average>=90 THEN grade='A';
ELSE IF Average<71 THEN grade='C';
  ELSE grade='B';
CARDS;
06101 88 79 98 06102 86 70 90 06103 60 70 90
06204 95 70 86 06205 82 92 76 06206 95 85 74
06307 68 98 78 06308 64 99 88 06309 87 78 88
;
TITLE '画出学生成绩表';
  PROC TABULATE DATA=score;
  KEYLABEL N='个案数' PCTN='个案的百分数';
CLASS CLASSid grade;
  VAR Average;
TABLE CLASSid, grade ALL, Average* (N* F=6. PCTN* F=10.2 MAX MIN MEAN)
  /RTS=15 BOX='平均成绩';
PROC PRINT;
RUN;
```

运行程序 5.19 后产生的部分输出见图 5.11。

5.7.7 用 PROC UNIVARIATE 过程做详尽的频数分布

1. 功能说明

UNIVARIATE 过程可对数字型变量进行描述统计。UNIVARIATE 过程不仅提供 MEAN、SUMMARY、TABULATE、FREQ 等过程所能产生的描述统计量,而且还输出变量的峰度、偏度、众数、中位数、四分位数等详细的描述统计量。同时还可输出以下的统计量:

- 输出与 FREQ 过程类似的频率表

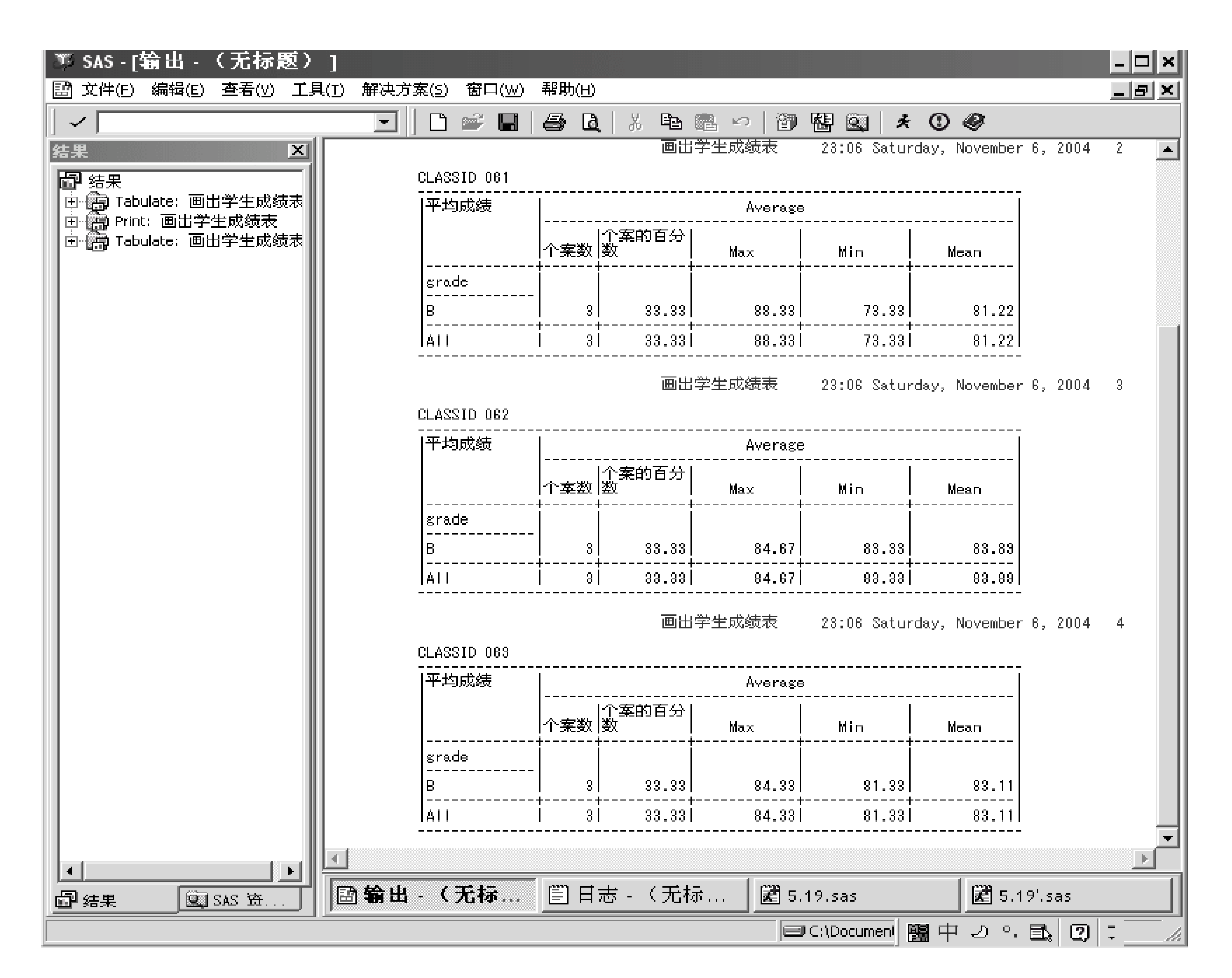


图 5.11 3 班学生成绩表

- 几个描述性分布图
- 数据分布的正态性检验
- 变量的极值

2. 命令格式

```
PROC UNIVARIATE 选项 (如 DATA= a);
     VAR 变量名; /* 例子见程序 5.20 * /
```

3. 选项

DATA＝a：数据集名称(如 a)如果省略，则调用最后建立的数据集。

FREQ：输出变量值、频数、百分比、累积百分比。

NORMAL：若输入的数据为正态分布，则输出检验的统计量。

PLOT：输出 3 幅图形，即茎叶图、BOX(盒)图、正态概率图。当定义诸如“BY sex;”语句时，则按 sex 变量值分组输出图形。

PCTPLOT＝值：此值等于 1～5，可指定这 5 种百分比计算中的一种，默认为 1。

VARDEF＝DF：用(自由度－1)值作为除数，默认值。

VARDEF＝WDF：用(权重和－1)值作为除数。

VARDEF＝N：用个案数 n 值作为除数。

VARDEF＝WGT(或 WEIGHT)：用权重和作为除数。

VAR v1：如不指定变量名(如 v1)，则输出全部变量的描述统计量。有 OUTPUT 语句就必须有 VAR 语句。

ID sex：以 sex 值的前 8 个字符写出 5 个最大值个案和 5 个最小值个案。ID 变量被写入任何的 OUTPUT 数据集里，而且 ID 变量值取自 PROC UNIVARIATE 过程中数据集的第 1 个个案，或“BY v;”变量组中的第 1 个个案。

OUTPUT OUT=数据集名 关键字=表 1 表 2：

对 OUTPUT 的说明：

(1) 关键字即统计量的名称。“表 1”名称对应 VAR 语句中的第 1 个变量，“表 2”名称对应 VAR 语句中的第 2 个变量，以此类推。

(2) 关键字所代表的统计量如下。

N：参加计算的个案数目。

NMISS：缺失的个案。

NOBS：个案数。

MEAN：均数(即均值)。

SUM：和数。

STD：标准偏差(即标准差)。

VAR：方差。

SKEWNESS：偏度。

KURTOSIS：峰度。

SUMWGT：权重和。

MAX：最大值。

MIN：最小值。

RANGE：全距。

Q1：下四分位数(25%位数)。

Q3：上四分位数(75%位数)。

MODE：众数。

MEDIAN：中位数。

Q RANGE：上下四分位数之差。即 Q3-Q1。

P1：第 1 百分位数。

P5：第 5 百分位数。

P10：第 10 百分位数。

P90：第 90 百分位数。

P95：第 95 百分位数。

P99：第 99 百分位数。

SIGNRANK：符号秩。

NORMAL：当 N≤2000 时计算 Shapiro-Wilk 的统计量。当 N>2000 时计算 Kolmogorov 的统计量。

程序 5.20：用 UNIVARIATE 统计血糖的频数分布细节。

```
DATA A1;
INPUT id sex age xt @@ ;
```

```
LABEL xt='血糖';
CARDS;
01 1 38 7.6 02 2 30 9.9 03 2 30 9.1 04 1 50 10.1
;
TITLE '用 UNIVARIATE 统计血糖的频数分布细节';
PROC UNIVARIATE DATA=A1 PLOT NORMAL FREQ VARDEF=N;
    VAR xt;
RUN;
```

运行程序 5.20 产生图 5.12 至图 5.15 所示的结果。

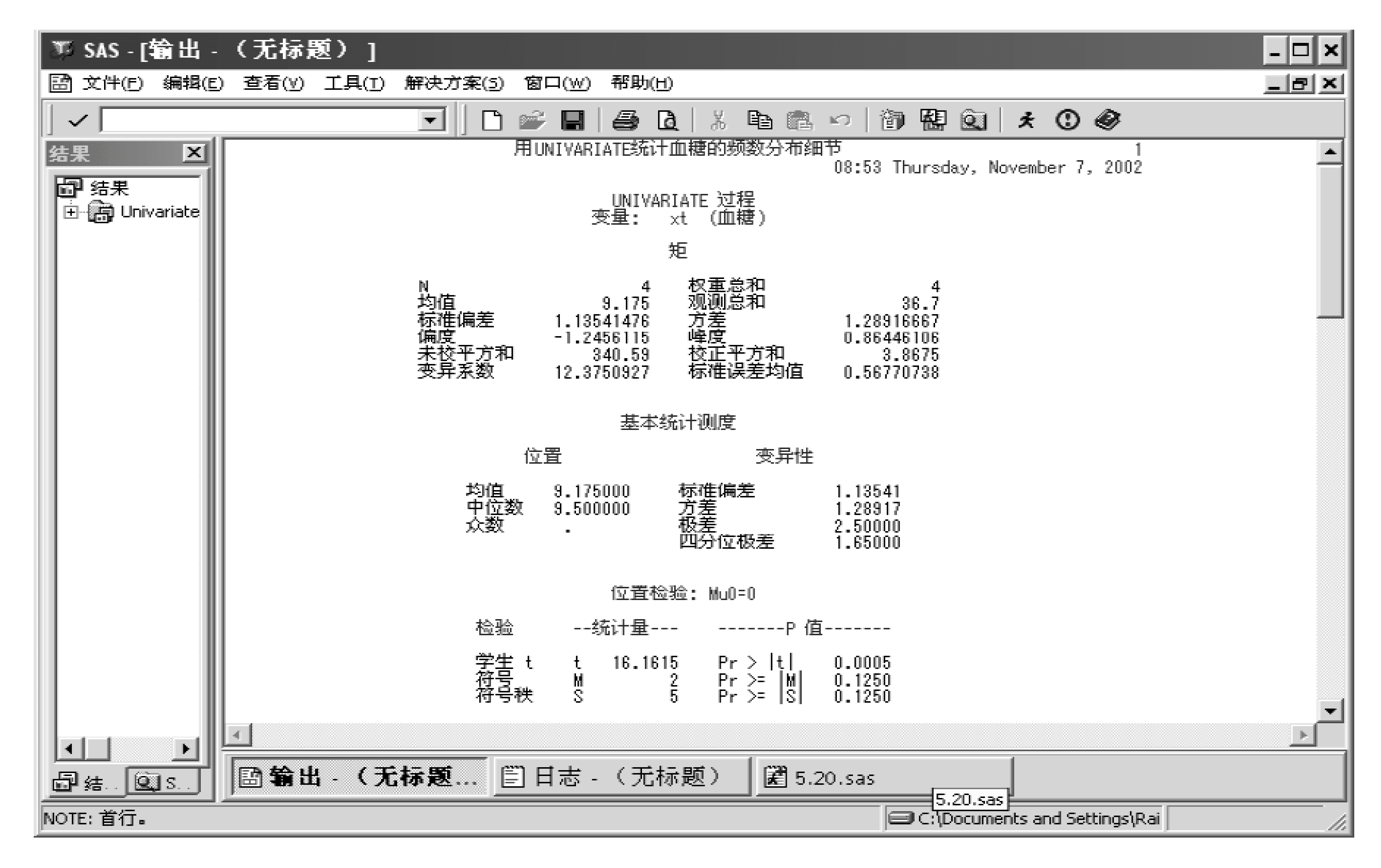


图 5.12 基本统计测度

从图 5.12 看,血糖均值为 9.175。标准偏差为 1.13541476,比较大。

从图 5.13 看,Pr<W 值为 0.3654,大于 α 值 0.05,没有理由拒绝"正态分布"的理论假设,所以数据呈现正态分布。

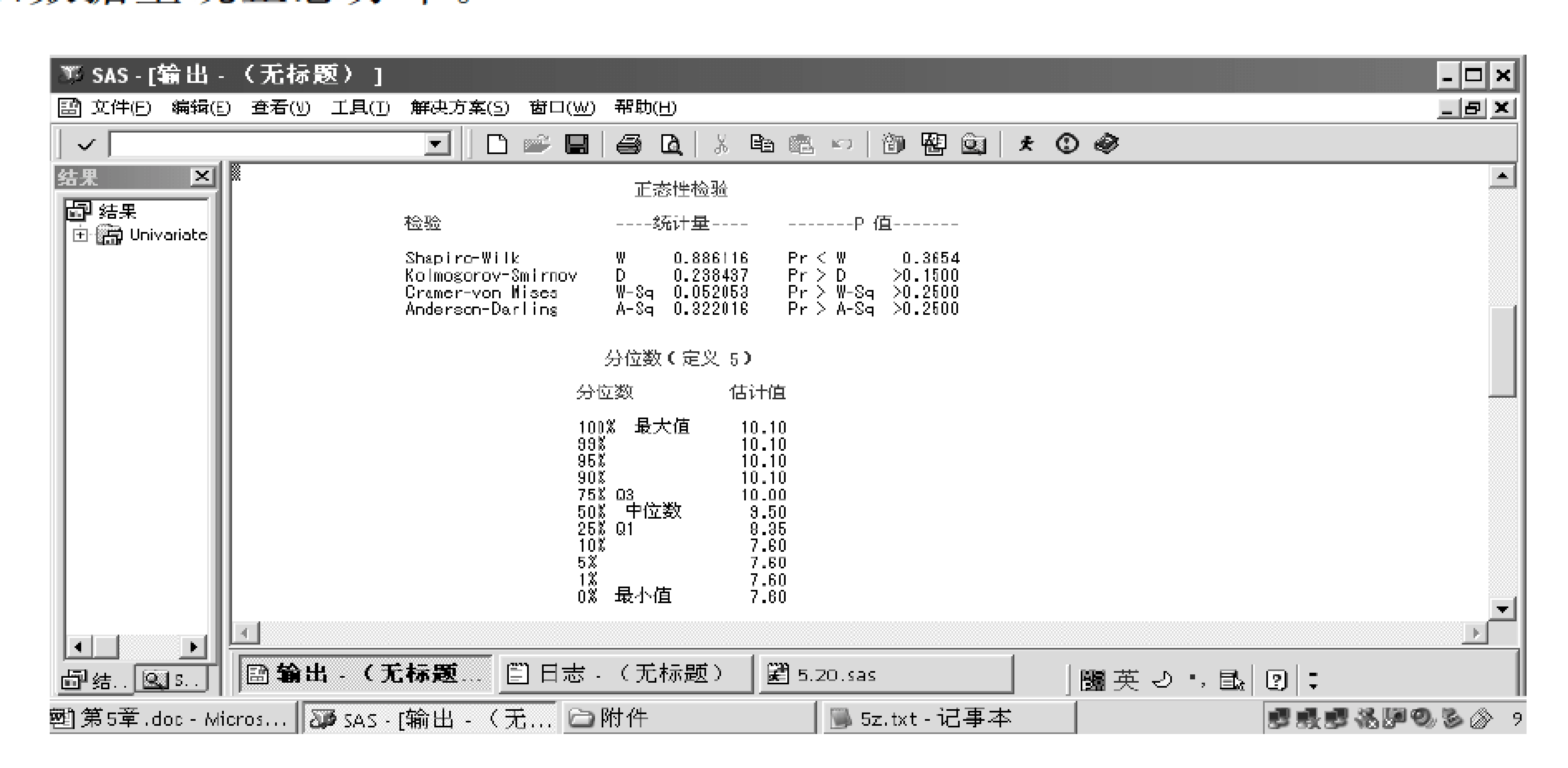


图 5.13 正态性检验

从图 5.14 的频数分布看，个案一共才 4 个人，不足以观察频数分布及 5 个最大值、5 个最小值。此处是讲解方法，实际上个案应该是成百上千个。

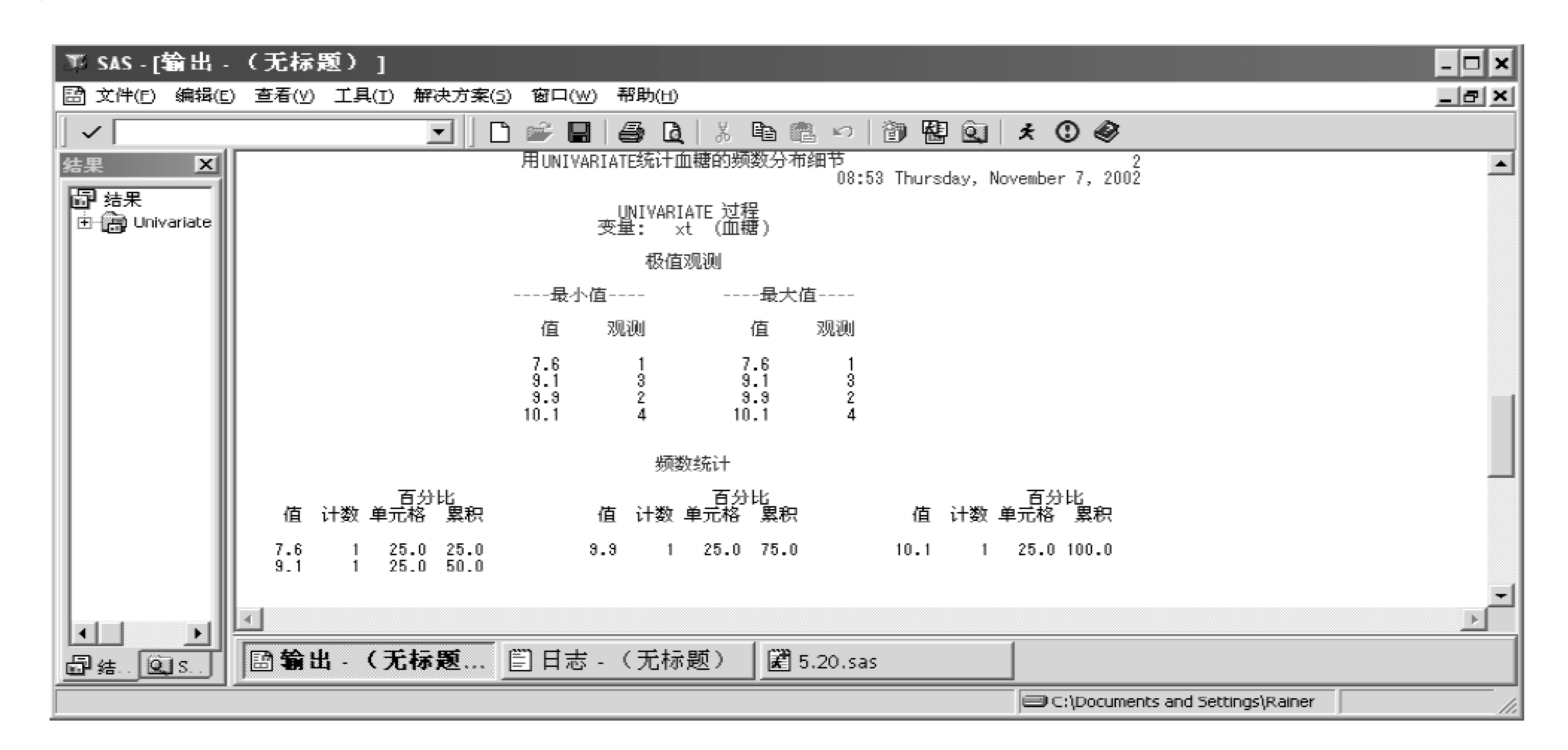


图 5.14 频数分布

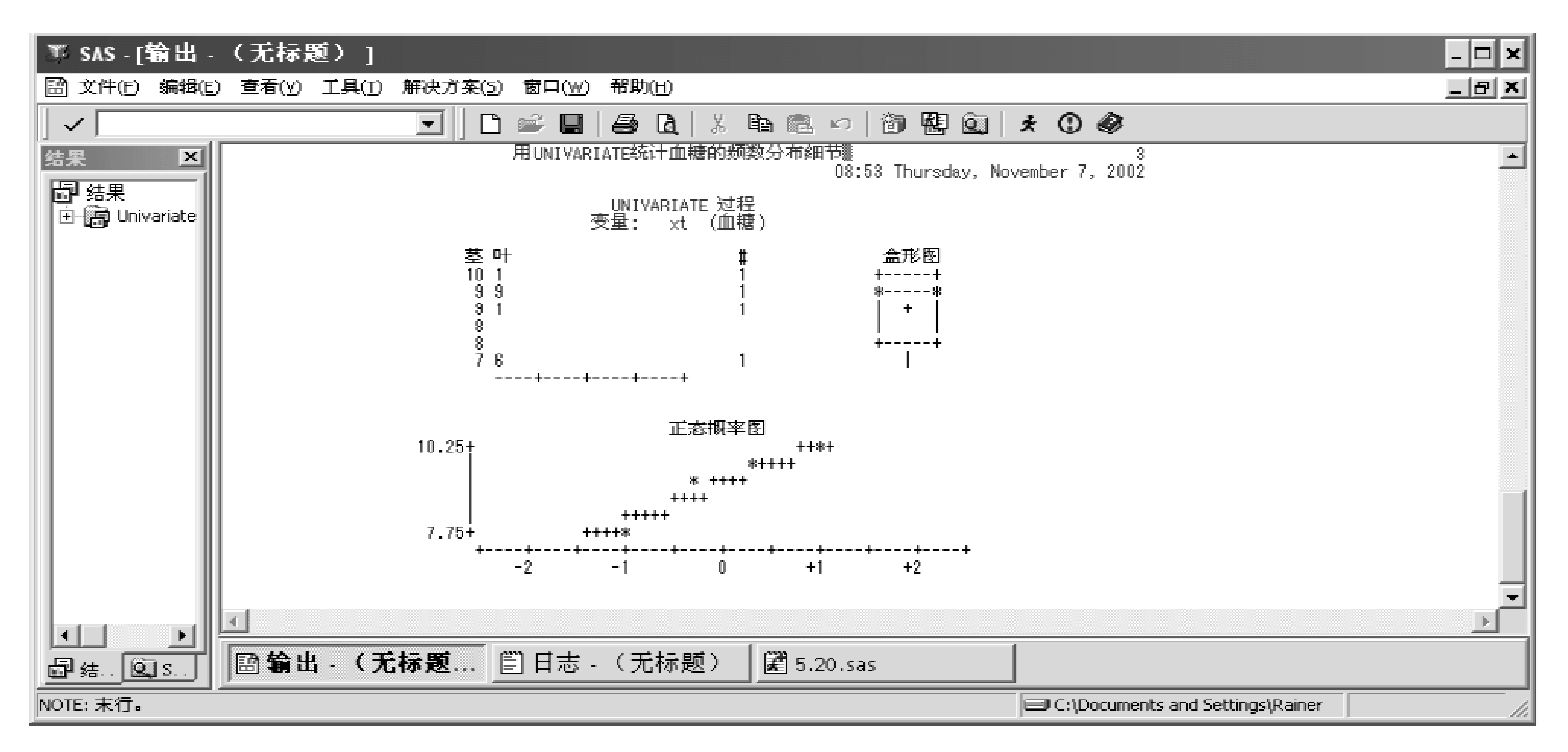


图 5.15 图形检验（盒图、正态概率图）

由于个案太少，所以茎叶图、盒图和正态概率图都未能画好。如果数据呈现正态性，正态概率图中的 * 点应该覆盖＋号。因为“ * ”号线表示数据实际的“分布线”，“＋”号线表示理论上要求的“正态线”。

5.7.8 用 PROC DBF 过程调用 dBASE 数据库数据

1. 命令格式

```
PROC DBF DB3=    DATA=    OUT=;
```

2. 格式说明

- DB3＝：调用 dBASEⅢ数据库数据进行 SAS 统计。如 PROC DBF DB3＝'四国.dbf'；
- DATA＝：把当前工作区中的数据集存储为 dBASEⅢ数据。如 PROC DATA＝'四国'。这时不能用“OUT＝”选项。
- OUT＝：把当前工作区中的 dBASEⅢ数据存储为 SAS 数据集。这时不能用“DATA＝”选项。

3. 用法说明

- 此过程只产生输出文件，不显示输出。
- dBASE 数据的后缀必须是“.DBF”。而且必须置于当前子目录下。
- dBASE 变量名长度为 10 个字符，转换为 SAS 数据集时自动被截成 8 个字符。
- dBASE 逻辑变量被变为长度为 1 列的 SAS 字符型变量。
- SAS 数据集里的变量转为 dBASEⅢ数据中的变量时，数字型的变量则变成了长度为 16 列的 dBASEⅢ字符型变量。
- SAS 数据集里的带小数点的变量转为 dBASEⅢ数据中的带小数点的变量时，必须用“FORMAT age4.1;”之类的格式将 SAS 变量联系在一起。否则 dBASEⅢ变量的小数点后面将无数字。

例 1：将当前子目录中的血糖数据转换为 dBASEⅢ类型的数据，见程序 5.21。

程序 5.21：

```
DATA A1;
INPUT id sex age xt @@ ;
LABEL xt='血糖';
CARDS;
01 1 38 7.6 02 2 30 9.9 03 2 30 9.1 04 1 50 10.1
;
TITLE '血糖数据由 SAS 数据集转为 dBASE 数据';
PROC DBF DB3=xt DATA=a1;
RUN;
```

运行程序 5.21 产生图 5.16 所示的结果。

例 2：将当前子目录中 dBASEⅢ类型的数据调入 SAS 系统进行处理，见程序 5.22。

程序 5.22：

```
DATA;
  PROC DBF DB3=xt OUT=xuet;
  PROC PRINT;
     VAR id xt;
  TITLE '血糖数据由 dBASE 数据转回 SAS 数据集';
```

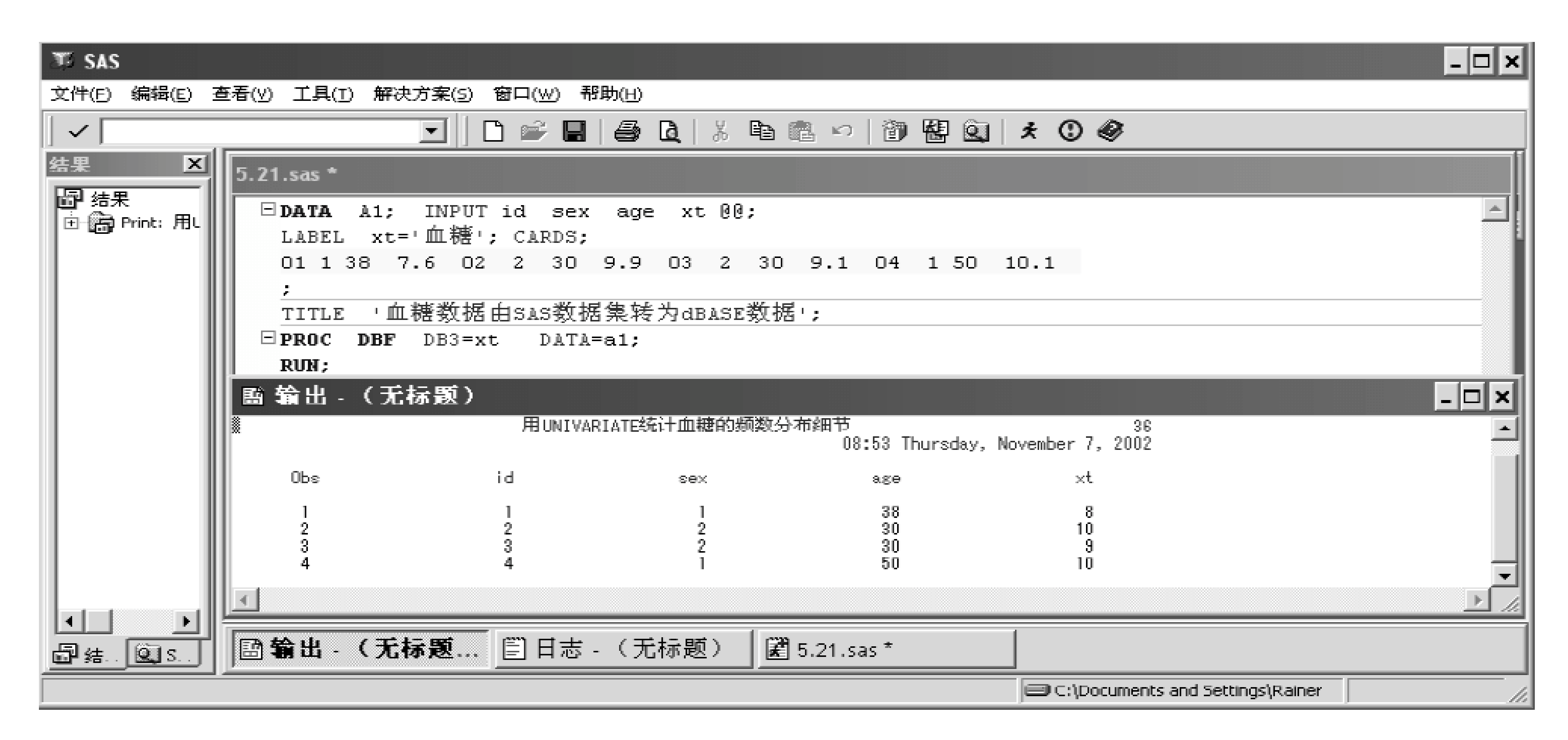


图 5.16 将血糖数据由 SAS 数据集转为 dBASE 数据

运行程序 5.22 产生图 5.17 所示的结果。

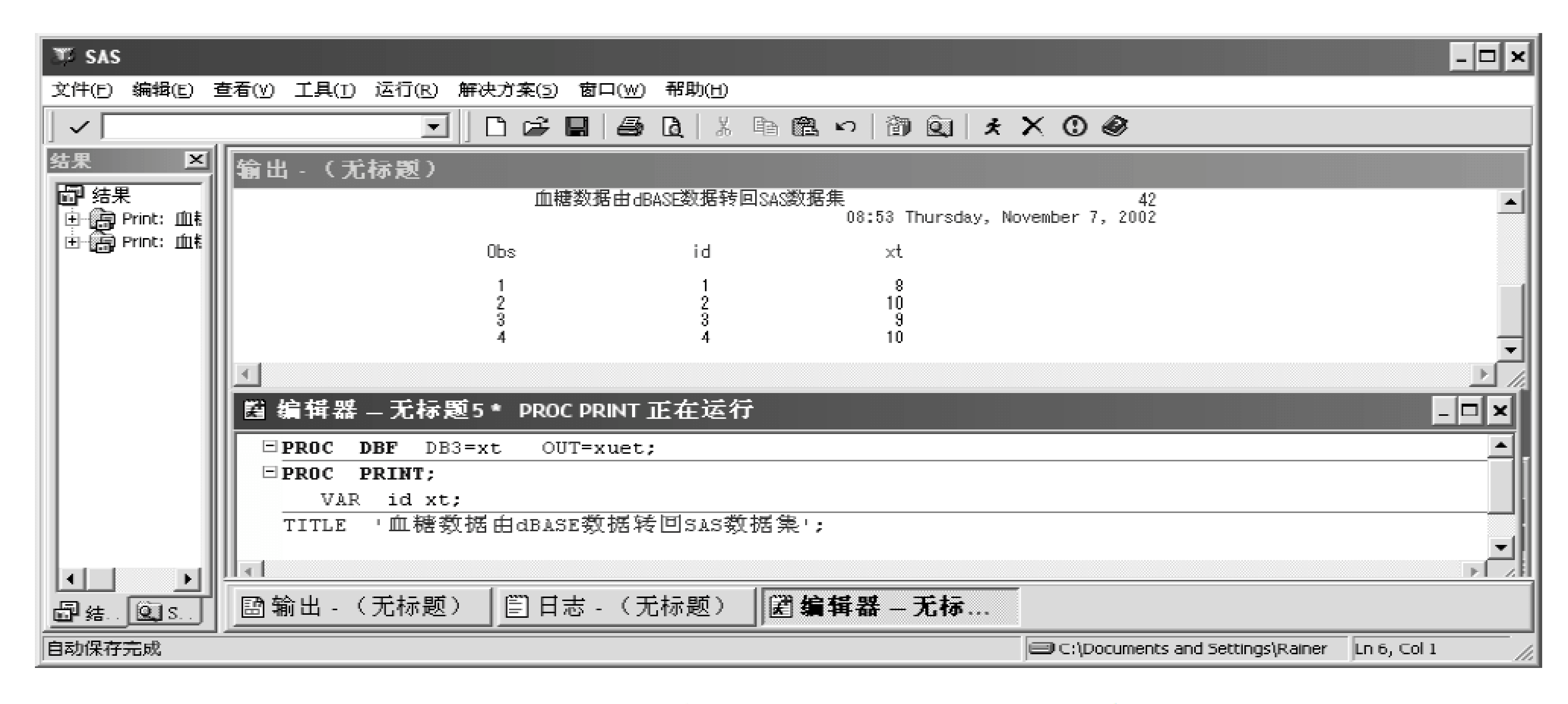


图 5.17 将 dBASEⅢ类型的数据调入 SAS 系统进行处理

5.7.9 用 PROC PRINT 过程显示数据集的信息

命令格式

PROC PRINT DATA=a; /* 显示 DATA 步指定过的数据集（例如数据集 a）的信息 */

VAR sex; /* 如果指定 VAR sex，则只显示变量 sex 值。如果不指定具体的变量名，则显示全部的变量值 */

或 PROC PRINT ID id; /* 用 id（个案号）代替默认的 OBS 序号，显示出图形的点 */

或 PROC PRINT ID id;BY sex; /* 按性别分组显示数据集的信息 */

或 PROC PRINT ID id; SUM inc1 inc2; /* 按收入分别求和之后显示每个和数 */

例 3：设 sex=1 为男性，sex=2 为女性。Location=1 为城市，Location=2 为农村。月收入变量为 inc1，月支出变量为 out。要求：

（1）按城乡显示男女两组的月总收入和月总支出。

（2）按性别与地区显示每人的月总收入和月总支出。

程序 5.23：

```
/* 程序 5.23* /
DATA p1;
INPUT id sex Location inc1 out1 @@ ;
CARDS;
01 1 1 3000 2500 02 2 2 4500 3800 03 2 1 3800 3700 04 1 2 8000 6000
;
PROC SORT; BY sex Location;
 PROC PRINT; BY sex Location;
 SUM inc1 out1;
RUN;
```

运行程序 5.23 产生图 5.18 所示的结果。

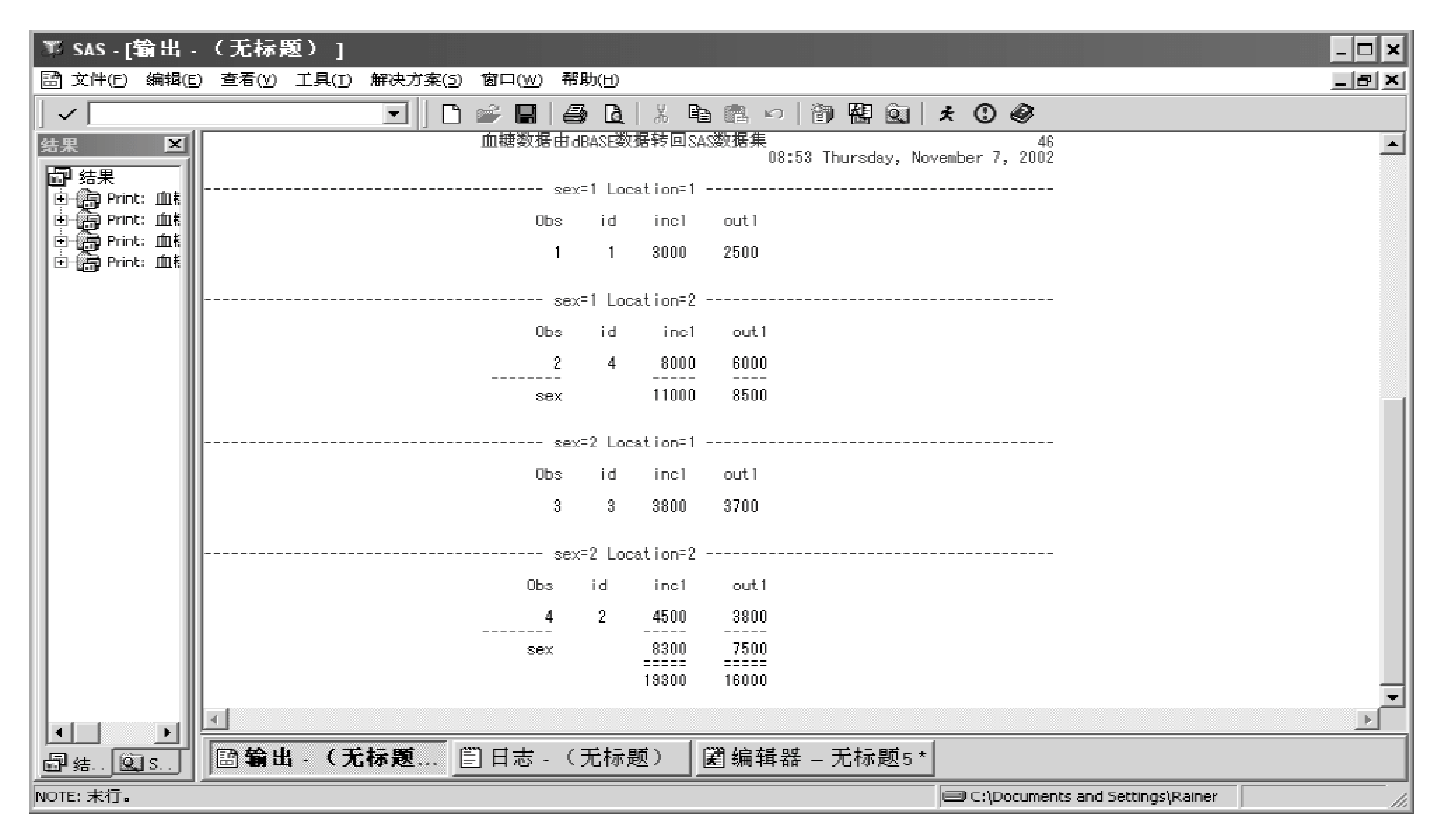


图 5.18　按城乡显示男女两组的月总收入和月总支出

从图 5.18 可以一目了然地看出：当 sex＝1 时，是第 1 个人和第 4 个人最先符合条件，所以月总收入为(3000＋8000)＝11000 元(见图 5.19)。

```
无标题 - 记事本
文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)
-------------------- sex=1 Location=1 ------------------------------
                 Obs   id   inc1   out1
                  1    1    3000   2500
-------------------- sex=1 Location=2 ------------------------------
                 Obs   id   inc1   out1
                  2    4    8000   6000
              --------      -----  ----
                 sex        11000  8500
```

图 5.19　当 sex＝1 时的个案月总收入

当 sex＝2 时，是第 3 个人和第 2 个人符合条件，所以总收入为(3800＋4500)＝8300 元，见图 5.20。

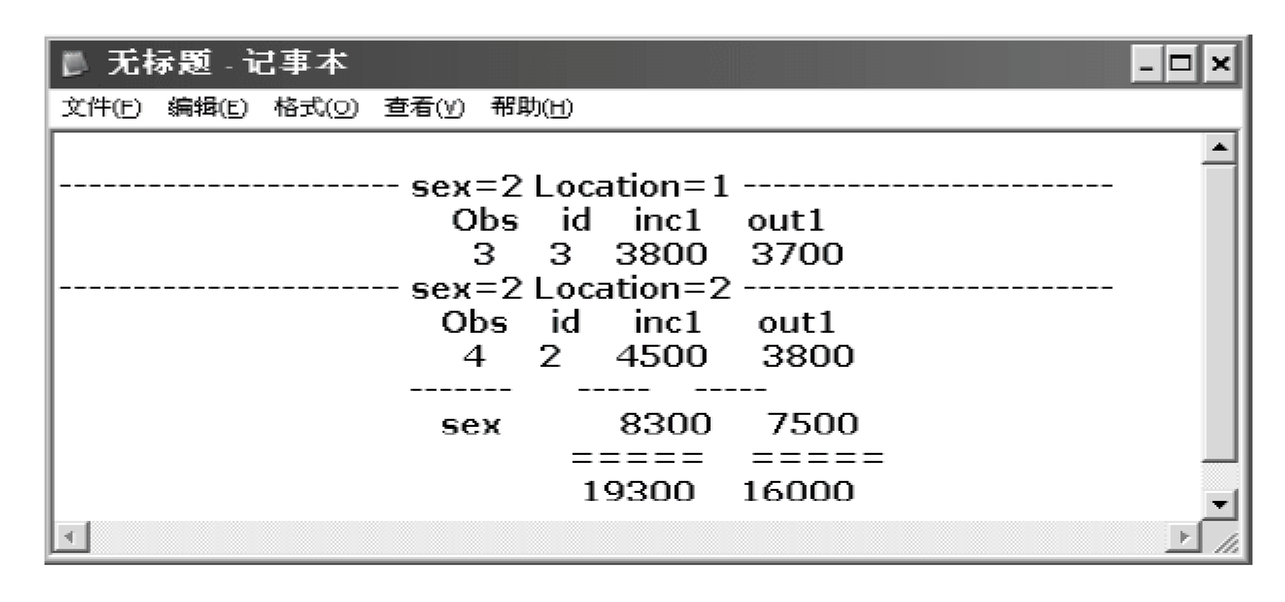


图 5.20 当 sex＝2 时的个案月总收入

其余以此类推。

5.7.10 用 PROC SORT 过程对数据排序

上面提到了排序问题，这里从统计过程的角度介绍它的过程命令及用法。

命令格式

```
PROC SORT 选项；/* 选项有 DATA= 名1  OUT= 名 2* /
    BY [DESCENDING] v1 v2;
```

例 4：

```
PROC SORT DATA= s2 OUT= sort1;
  BY DESCENDING sex age;/* 先按变量 sex 值降序排序个案,sex 值相同的个案再按 age 值升序排
                         序 * /
```

程序 5.24：

```
/* 程序 5.24* /
DATA s2 ;
INPUT id sex $  location score1-score3 @ @  ;
    s= SUM(OF score1-score3);
Average= MEAN(OF score1-score3);
CARDS;
001 m 1 80 90 88 002 f 2 78 89 91 003 m 2 82 93 90 004 f 1 90 87 89
;
TITLE '按性别地区升序排序学生成绩表';
PROC SORT DATA= s2;
  BY sex location;
PROC PRINT;
RUN;
```

运行程序 5.24 产生图 5.21 所示的结果。

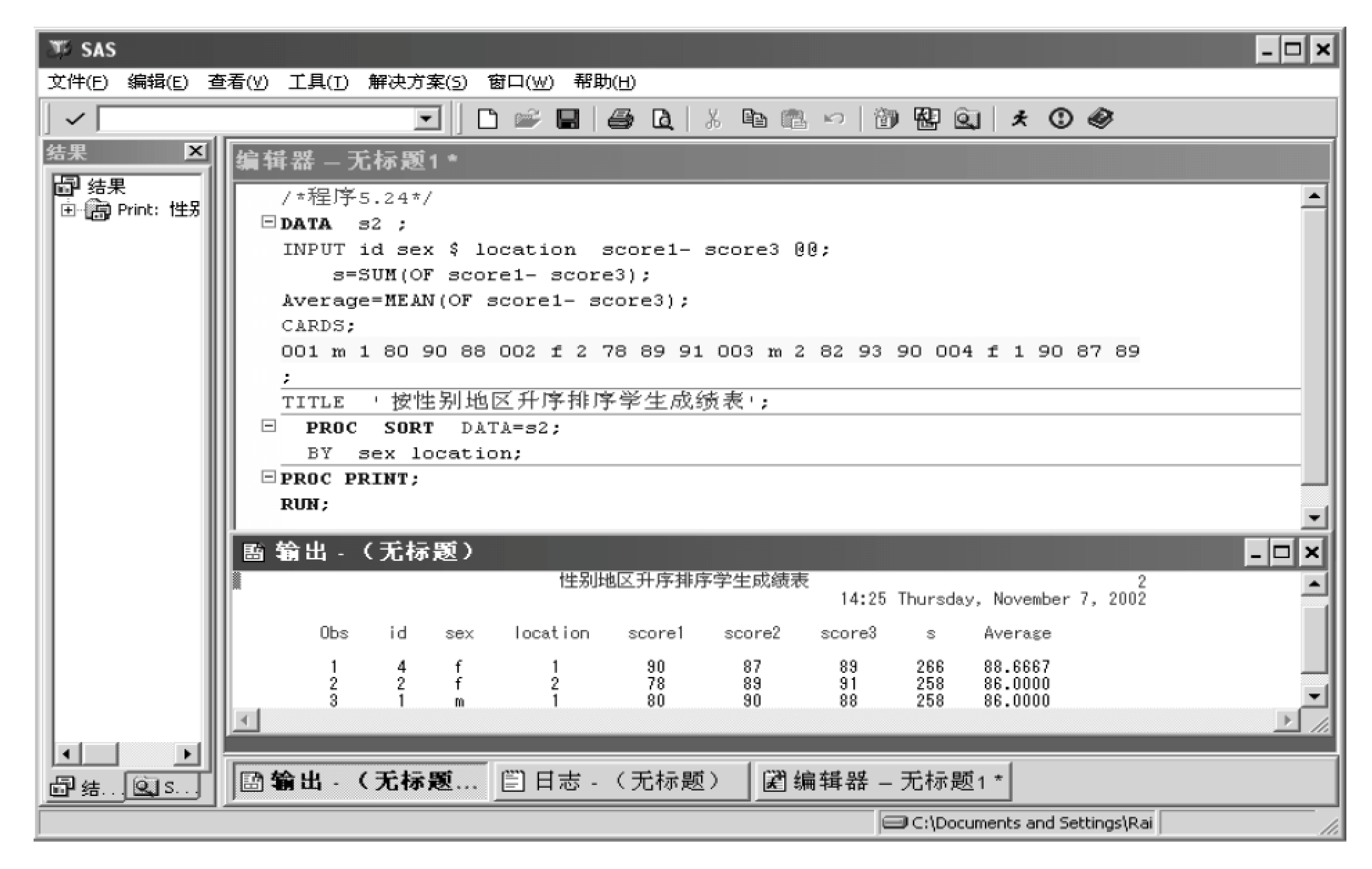


图 5.21　按性别地区升序排序学生成绩表

5.7.11　用 PROC STANDARD 过程对变量标准化

标准化变量是把变量标准化成均值为某值(如 80),标准偏差为某值(如 5)。或把变量标准化成均值为 0,标准偏差为 1 的正态性变量。

1. 命令格式

```
PROC STANDARD 选项; /* 选项见 2.* /
    VAR v1 v2;
```

2. 选项

DATA＝v1:如果默认数据集名称“v1”,则对工作区里的数据集的变量进行标准化。

OUT＝o1:如果默认数据集名称“o1”,则标准化后将结果存入 DATA1。

VARDEF＝DF:用(自由度－1)值作为除数,默认值。

VARDEF＝WDF:用(权重和－1)值作为除数。

VARDEF＝N:用个案数 n 值作为除数。

VARDEF＝WGT(或 WEIGHT):用“权重和”作为除数。

MEAN＝:计算均值,如 MEAN＝80。

STD＝:标准偏差,如 STD＝5。

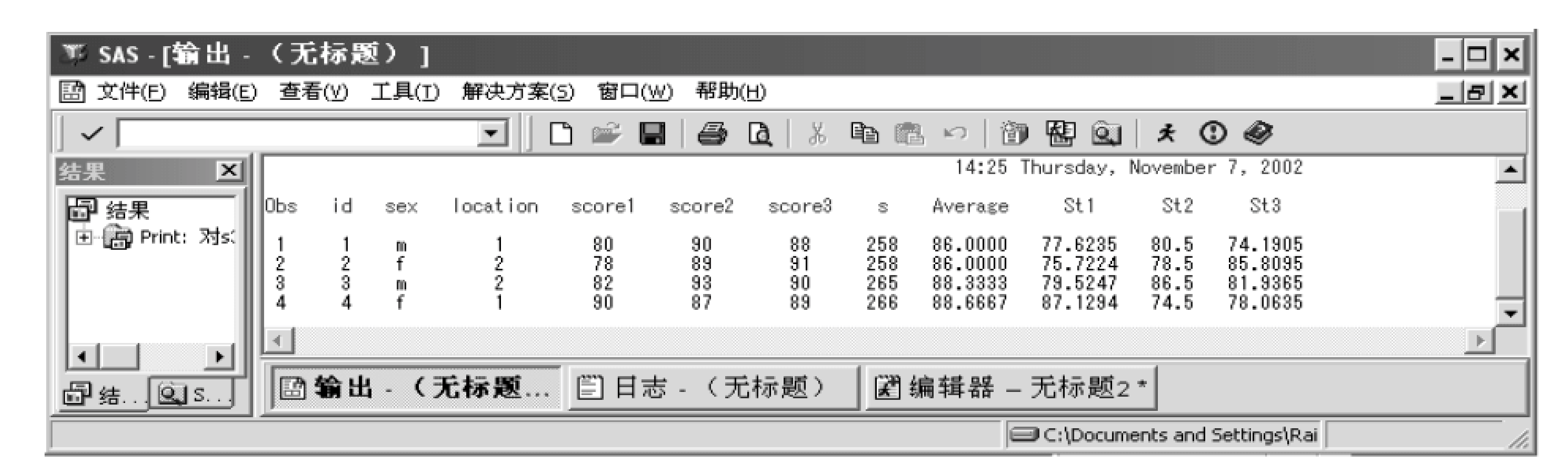
REPLACE:用均值取代缺失值。

程序 5.25:

```
DATA s2 ;
```

```
INPUT id sex $  location score1-score3 @@ ;
     s=SUM(OF score1-score3);
Average=MEAN(OF score1-score3);
CARDS;
001 m 1 80 90 88 002 f 2 78 89 91 003 m 2 82 93 90 004 f 1 90 87 89
;
DATA s3;
SET s2;
St1=score1; St2=score2; St3=score3;
PROC STANDARD DATA=s3 OUT=scorestd M=80 STD=5 REPLACE;
     VAR St1-St3;
PROC PRINT DATA=scorestd;
TITLE '对 s3数据集里的变量进行标准化';
PROC MEANS DATA=scorestd MAX=2 MEAN STD;
TITLE '显示数据集里的变量的均值';
```

运行程序 5.25 产生图 5.22 和图 5.23 所示的结果。


14:25 Thursday, November 7, 2002

Obs	id	sex	location	score1	score2	score3	s	Average	St1	St2	St3
1	1	m	1	80	90	88	258	86.0000	77.6235	80.5	74.1905
2	2	f	2	78	89	91	258	86.0000	75.7224	78.5	85.8095
3	3	m	2	82	93	90	265	88.3333	79.5247	86.5	81.9365
4	4	f	1	90	87	89	266	88.6667	87.1294	74.5	78.0635

图 5.22 标准化后的变量

SAS - [输出 - （无标题）]

文件(F) 编辑(E) 查看(V) 工具(T) 解决方案(S) 窗口(W) 帮助(H)

对s3数据集中的变量进行标准化 18

14:25 Thursday, November 7, 2002

Obs	id	sex	location	score1	score2	score3	s	Average	St1	St2	St3
1	1	m	1	80	90	88	258	86.0000	77.6235	80.5	73.3185
2	2	f	2	78	89	91	258	86.0000	75.7224	78.5	85.3452
3	3	m	2	82	93	90	265	88.3333	79.5247	86.5	81.3363
4	4	f	1	90	87	.	177	88.5000	87.1294	74.5	80.0000

ros... SAS - [输出 - （无... 第5章.doc - Micros... 5z.txt - 记事本

图 5.23 st3 变量的缺失值用 80 替代

从图 5.22 可以看出，St1 值是对原变量 score1 的标准化值。St2 值是对原变量 score2 的标准化值。St3 值是对原变量 score3 的标准化值。图 5.22 还显示每个人每次的平均成绩(见 Average 值)。

图 5.23 中，当原始变量 score3 有缺失值时，就用已指定的均分(如 80 分)替代，请上机实习。

5.7.12 用 TRANSPOSE 过程转置数据

本节用 TRANSPOSE 过程将数据集里的数据进行行列转置，即每行的个案号变成列变量，原来的列变量变成行变量。

命令格式：

```
PROC TRANSPOSE;
    VAR v1 v2; /* 若不指定变量名,则对所有的变量转置 */
```

例如，要将表 5.1 的数据变成表 5.2 所示。

表 5.1　转置前的原始数据

OBS	A	B	C
1	20	30	40
2	25	35	45

表 5.2　转置后的数据

OBS	_NAME_	COL1	COL2
1	A	20	25
2	B	30	35
3	C	40	45

为此，编辑出程序 5.26 的命令语句。

程序 5.26：

```
DATA t1;
INPUT A B C;
CARDS;
20 30 40 25 35 45
;
PROC PRINT;
TITLE '转置前的原始数据';
PROC TRANSPOSE; /* 行变量变成列变量,原来的列变量变成行变量 */
PROC PRINT;
TITLE '转置后的数据';
RUN;
```

运行程序 5.26 产生图 5.24 所示的结果。

从图 5.24 的“日志”窗口看，行列转换成功，结果显示在输出窗口。

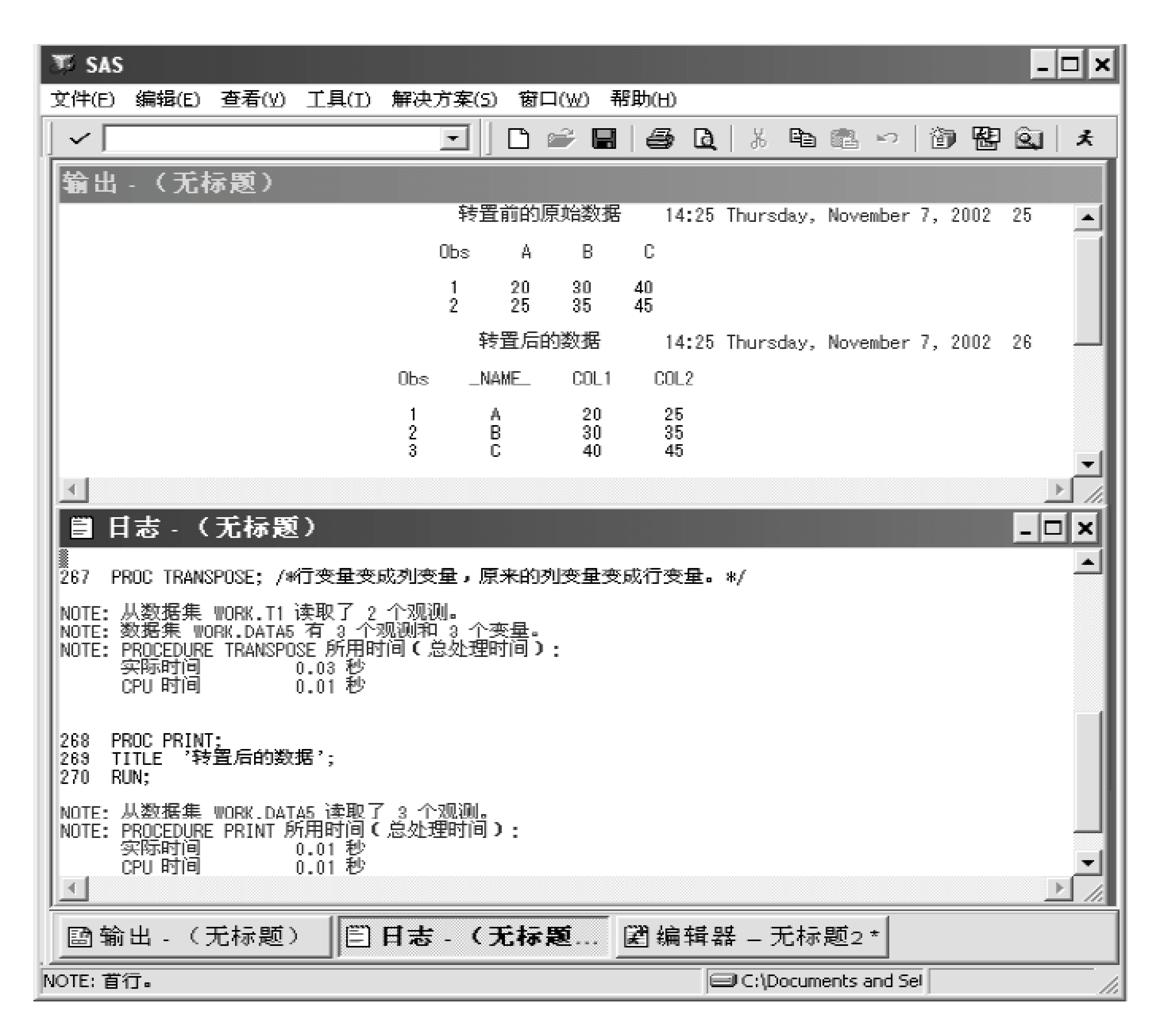


图 5.24 数据的行列转置

习 题 5

1. 试举一个用“INFILE 语句”读取 D:\my.txt 纯文本的数据文件的例子。
2. 试举一个用“FORMAT 及 VALUE 语句”定义数值标签的数据文件的例子。
3. 试举一个数据重新编码的例子。
4. 试将输入的日期输出为日-月-年格式。
5. 试画出 SEX 变量的条形图。
6. 试画出 SEX 变量的圆形图。
7. 试画出 AGE 变量的直方图。
8. 试用 PROC MEANS 过程统计年龄的均值分布。
9. 用 PROC RANK 过程统计年龄的秩和分布。
10. 试用 PROC TABULATE 制作简单的成绩表。
11. 试用 PROC UNIVARIATE 过程统计血糖的详细频数分布。

第 6 章 chapter 6

描 述 统 计

数据挖掘的第一步是计算变量值的频数分布,从中可以看到数据的特性。具体分为单变量的频数统计和双变量交叉汇总的频数统计。PROC FREQ 过程可以做单变量的频率表和双变量的频率表。

通过频数统计不仅可以看出数据分布,而且还可以检查数据输入正确与否。如当检测到男性堕胎 1 次,则说明数据有误。

6.1 用 FREQ 过程做单双变量的频数统计

6.1.1 FREQ 过程命令

在 FREQ 过程中使用的命令语句如下:

```
PROC FREQ [选项 1] ;
TABLES 变量 1 变量 2 [选项 2] ;
TABLES 写法;
TABLES A-C ;            /* 单变量的频数统计。相当于 TABLES A B C ; */
TABLES (A-B) * C ;      /* 双变量的频数统计。相当于 TABLES A* C B* C ; */
TABLES A* (B-C );       /* 双变量的频数统计。相当于 TABLES A* B A* C ; */
TABLES (A B C) * D ;    /* 双变量的频数统计。相当于 TABLES A* D B* D C* D; */
TABLES (A B) (C* D) ;   /* 双变量的频数统计。相当于 TABLES A* C A* D B* C B* D ; */
```

若无 TABLES 语句,则对 FREQ 过程中的全部变量进行频数统计。

```
WEIGHT v1;              /* 用变量 v1 的值加权 */
BY v2;                  /* 按变量 v2 的值分组统计 */
```

1. [选项 1]的内容(任选 1 项)

```
DATA=数据集 1; /* 若省略数据集 1 等名称,则使用工作区中新建的数据集做统计 */
```

```
ORDER= DATA;     /* 按数据集里的个案顺序显示 */
ORDER= FREQ;     /* 按频数递减地显示频率表,最大频数在前 */
ORDER= INTERVAL;
```

```
ORDER= FORMATTED;
```

例 1：PROC FREQ DATA＝score ORDER＝FREQ；

2. ［选项 2］的内容（任选 1 项）

（1）统计量（任选 1 项）

```
TABLES sex [/EXACT];  /* 显示大于 2* 2表格的 Fisher精确检验 * /
TABLES sex /CHISQ;    /* 显示卡方检验及基于卡方检验的泊松卡方、似然比卡方、曼特尔-亨撒尔
                         (Mantel-Haenszl)卡方,以及 Phi 系数、列联系数、Cramer 的 V、2* 2表格的
                         Fisher精确检验 * /
```

此外，还有 CMH、ALL、Measures 等选项。

（2）显示统计量

```
TABLES sex [/EXPECTED];   /* 显示期望频数 * /
           /DEVIATION;    /* 显示偏差 * /
           /CELLCHI2;     /* 显示每个单元对总体卡方的贡献 * /
           /CUMCOL;       /* 显示列累积百分比 * /
           /MISSIPRINT;   /* 显示缺失值的频数 * /
           /SPARSE;       /* 显示 TABLES语句中的变量 * /
           /MISSING;      /* 把缺失值当有效值统计 * /
           /LIST ;        /* 慎用。它只以一般的列表方式显示频数 * /
           /OUT=数据集;/* 若 TABLES语句中有多个变量表,则存储最后一个变量表里的变量值和
                          频数 * /
```

（3）不显示统计量

```
TABLES sex /NOPRINT;     /* 取消表格。不显示 TABLES语句中所有变量的频率表 * /
TABLES sex [/NOFREQ];    /* 不显示频数 * /
           /NOPERCENT;   /* 不显示百分比 * /
           /NOROW;       /* 不显示行百分比 * /
           /NOCOL;       /* 不显示列百分比 * /
           /NOCUM;       /* 不显示一维频数、累积频数和百分比 * /
```

例 2：产生二维表格。

```
PROC FREQ DATA= score ORDER= FREQ ;
    TABLES sex* edu /MISSING CHISQ ; /* 第 1个变量 sex是行变量,第 2个变量 edu是列变量 * /
```

例 3：产生三维表格。

```
PROC FREQ DATA= score ORDER= FREQ ;
  TABLES location sex* edu /MISSING CHISQ ; /* 第 1个变量 location是控制两个子表的变量,第 2个
                                               变量 sex是行变量,第 3个变量 edu是列变量 * /
```

6.1.2 FREQ过程与其他过程的连用

PROC FREQ 过程不仅能显示单变量的频数、百分比，而且可计算和显示双变量或

多变量的频数、百分比、卡方检验、相关系数、期望频数等。

其他过程如 CHART 可产生频数和条形图，SUMMARAY 可产生频数和数据集，TABULATE 可产生频数和表格。

6.2 单变量频数分布

单变量频数分布一般是对标称数据(NOMINAL DATA，也称定类型数据、名义数据)、次序数据(ORDERNAL DATA，也称定序型数据)计算频数。如性别(sex)、文化水平(edu)、地区(location)、种族(race)等是标称数据。奖金等级、年龄组、分数段、工资档次等是次序变量。

例 4：计算性别、文化水平变量的一维频数分布，见程序 6.1a。

程序 6.1a：

```
LIBNAME LB 'D:\';
DATA LB.xt2;
LABEL edu='文化水平' sex='性别' l='含磷' xt='血糖';
INPUT id sex edu xt l @@ ;
CARDS;
001 1 1 8.1 3.1 002 2 2 9.1 2.8
003 1 3 9.0 4.8 004 2 3 8.7 5.1
005 1 2 . 4.7 006 2 . 6.2 .
PROC FREQ DATA=LB.xt2;
TABLES sex edu;
PROC PRINT;
RUN;
```

运行程序 6.1a 可产生图 6.1(a)所示的结果。

1. 频率表解释

频数：频次。出现的次数。

百分比：频数/总数 * 100%。

累积频数：各组频数的累积。

累积百分比：上下各组百分比的累积。

频数缺失=1：说明有 1 人的数据为缺失值，而且没有参与计算。

2. 分析图 6.1 的频率表

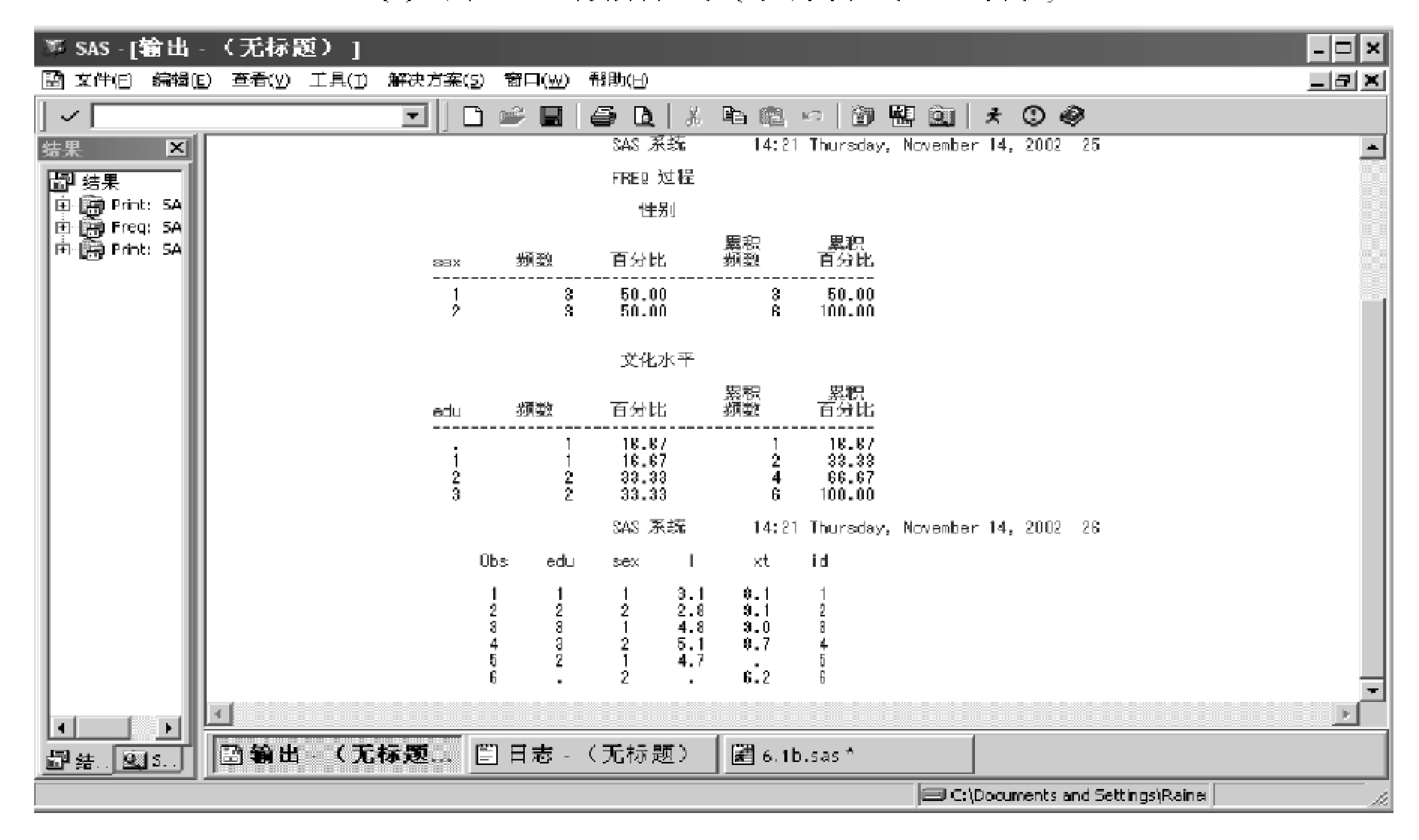
图 6.1(a)是性别和文化水平的单变量频率表，从性别的频率表看，sex=1 一组频数为 3 人，占总人数 6 人的 50%。sex=2 一组频数为 3 人，占总人数 6 人的 50%。

3. 把缺失值当有效值统计

程序 6.1b：



(a) 单变量的频数统计（不含缺失值的计算）



(b) 单变量的频数统计（含缺失值的计算）

图 6.1　单变量的频数统计结果

```
LIBNAME LB 'D:\';
DATA LB.xt2;
LABEL edu= '文化水平' sex= '性别' l= '含磷' xt= '血糖';
INPUT id sex edu xt l @@ ;
CARDS;
001 1 1 8.1 3.1 002 2 2 9.1 2.8
003 1 3 9.0 4.8 004 2 3 8.7 5.1
005 1 2 . 4.7 006 2 . 6.2 .
PROC FREQ DATA= LB.xt2;
```

```
TABLES sex edu /MISSING;/* 把缺失值当有效值统计 * /
PROC PRINT;
RUN;
```

运行程序 6.1b 可产生图 6.1(b)所示的结果。

图 6.1(b)是性别和文化水平的单变量频率表,但缺失值 1 人也参与频数和百分比统计。从文化水平的频率表看,sex=2 一组有 1 人的文化水平没有填写而作为缺失值计算,它占总人数 6 人的 16.67%,是不小的比例。所以在计算时要考虑缺失值的排除。

6.3 双变量交叉汇总和结合测量

在实际工作中,更多的是统计比较复杂的问题诸如性别与文化水平、性别与收入、年龄与收入之类双变量交叉发生的频数问题,以及双变量结合的强度问题。这时应该考虑用 PROC FREQ 过程。

用 PROC FREQ 过程进行双变量交叉时还可以对温度、气温两个区间数据(Interval data)和年龄、收入、成绩、血压、体重等比例数据(Ratio data)进行均值、标准偏差、全距、峰度、偏度等测量。

6.3.1 双变量频数统计的过程命令

双变量频数统计的过程命令语句如下:

```
PROC FREQ DATA= 输入数据集;
  TABLE A* (B C)/CHISQ EXACT;
或 TABLE 页 * A* B* C/CHISQ EXACT;
即 TABLE 页 * 行 * 列 /CHISQ EXACT;
或 TABLE 行 * 列 /CHISQ ;
```

关于 FREQ 过程的详细命令及用法详见 6.1.1 节。

6.3.2 “定类-定类”双变量交叉汇总与结合测量

例 5:FREQ 主命令和 TABLE 子命令及选项的应用简例。

程序 6.2:“定类-定类”双变量交叉汇总与结合测量。

```
DATA;
DATA f1;
INPUT sex edu @@ ;
CARDS;
1 4 2 3 1 2 2 2 1 1 2 2 . 3 2 .
;
PROC FREQ ORDER= FREQ;
TABLES sex* edu/CHISQ EXACT ; /* 计算卡方分布及费歇尔的精确检验 * /
TITLE '两维频率表,按频率值降序排序';
```

```
RUN;
```

运行程序 6.2 产生图 6.2 和图 6.3 所示的结果。

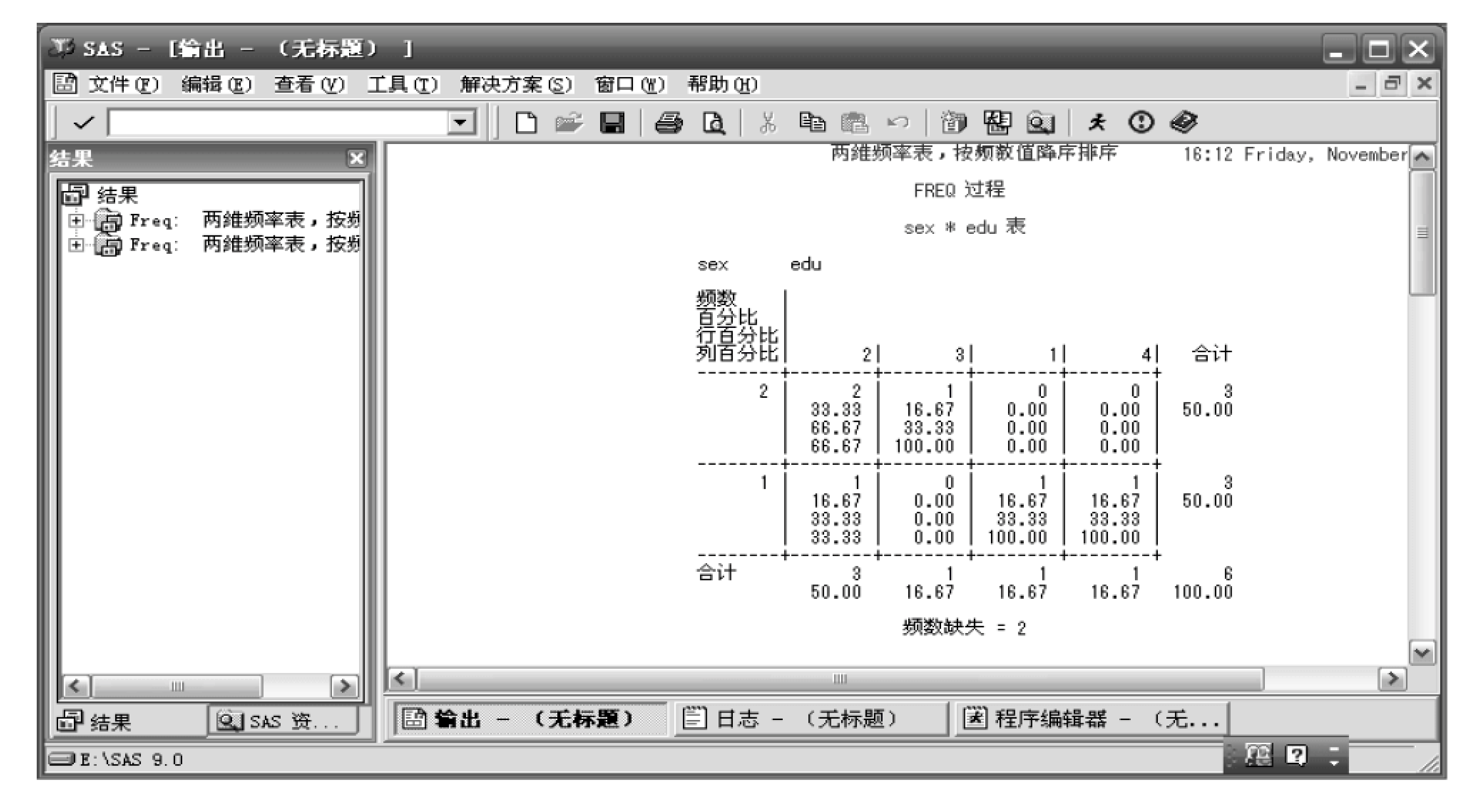


图 6.2　由 TABEL sex * edu 产生的交叉汇总表

1. 单元内容分析

如何正确观察交叉汇总表中的百分比，是关系到会不会分析比较双变量之间关系的技术问题。

比较的准则是：如果自变量在行上则看行百分比，自变量在列上则看列百分比。

如图 6.2 所示，性别是自变量，且在行上，因此要看行百分比。从图 6.2 看：性别为 2(女性)，文化水平为 2(大专文化)的有 2 人，占女性总体 3 人中的 33.33%(行百分比)，比例最大。说明在女性中，大专文化水平的人最多。

其他分析以此类推。

2. 双变量结合程度

从图 6.3 看，由于 sex * edu 属于"定类-定比"类型的数据，但为了说明统计分析的方法暂认为是"定类-定类"类型的数据，即看莎姆斯的 D 系数或 Φ 系数。因为系数都比较大(大于 0.3)，说明两个样本呈现相关。

3. 总体推论

H0：行列变量互为独立。

双变量结合测量的核心是卡方检验。它检验"行列变量互为独立"的总体推论。

检验：从图 6.3 看，样本小，而且卡方 3.3333，自由度 3，计算后的显著性水平为 0.3430，0.3430大于 α 值 0.05，所以没有足够的理由拒绝 H0，即总体上说，sex 和 edu 双变量互为独立。

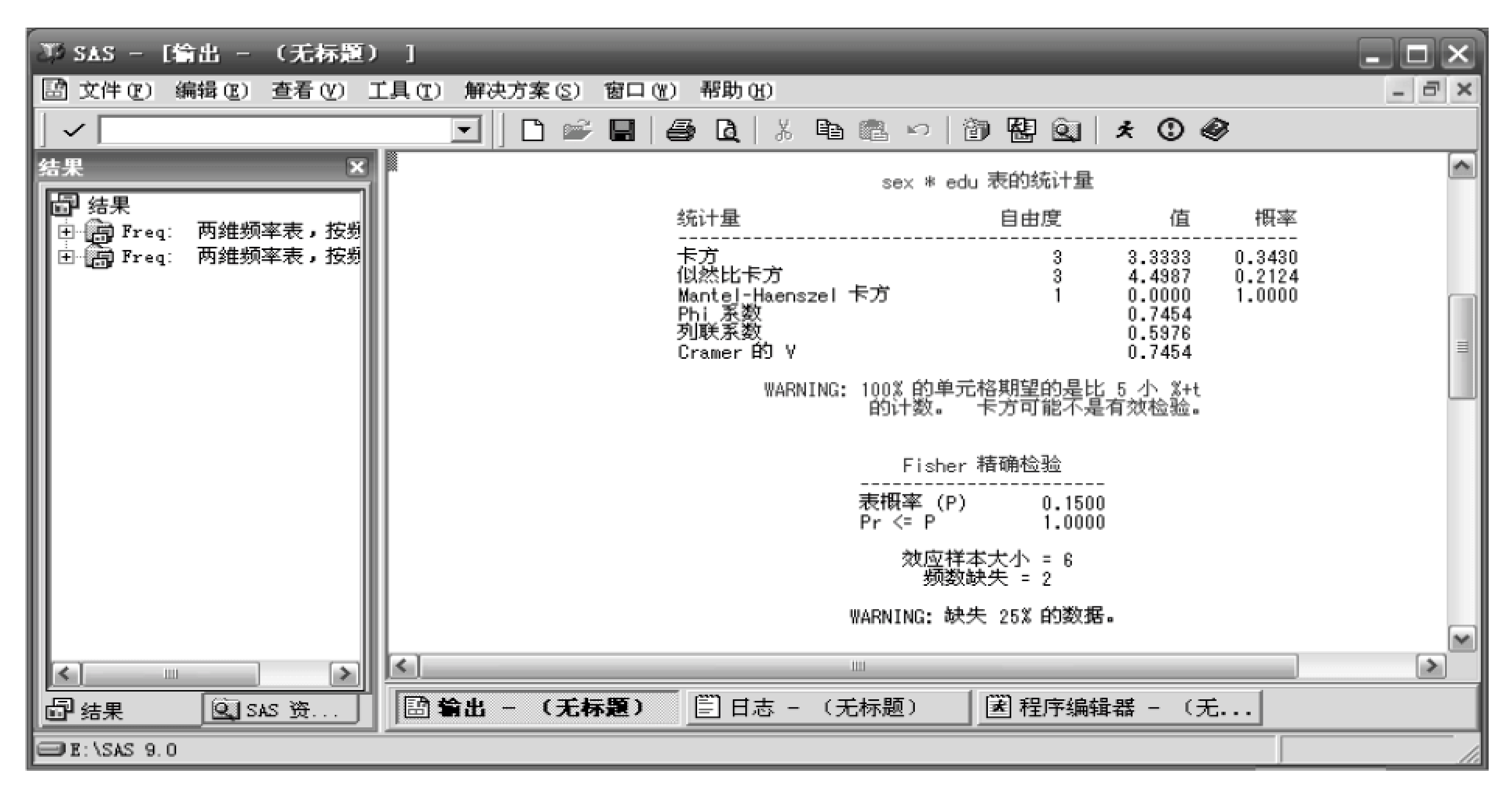


图 6.3　由 TABEL sex * edu 产生的结合测量

6.3.3 “定比-定比”双变量交叉汇总与结合测量

例 6：程序 6.3 中的数据是高血压、脉搏、血糖、日抽烟量及性别方面的数据，试计算日抽烟量与高血压的关系（分组后成为“定序-定序”数据）、高血压与血糖的关系（定比-定比）、性别与血糖的关系（定类-定比）。

程序 6.3：“定比-定比”变量的测量。

```
DATA xt;
INPUT sex location cy dy mp xt 3.1;
LABEL location='地区' sex='性别' cy='抽烟量:支'
 mp='脉搏' dy='低压' xt='血糖';
CARDS;
1 1 15 85 66 8.5
2 1 20 88 68 7.3
1 2 30 90 70 8.6
2 2 25 90 70 7.8
1 1 35 95 75 9.2
2 2 35 91 76 9.0
;
PROC FORMAT;
  VALUE cyF LOW-10=1 11-20=2 21-HIGH=3;
  VALUE dyF LOW-85=1 86-90=2 91-HIGH=3;
  FORMAT cy cyF. Dy dyF.;
PROC FREQ; /* 产生图 6.4 * /
TABLE mp * xt/ ALL;
```

运行程序 6.3 产生图 6.4 所示的结果。

结果分析：

1. 双变量结合程度

从图 6.4 看，由于 mp * xt 属于“定比-定比”类型的数据，应该观察 CORR 系数或

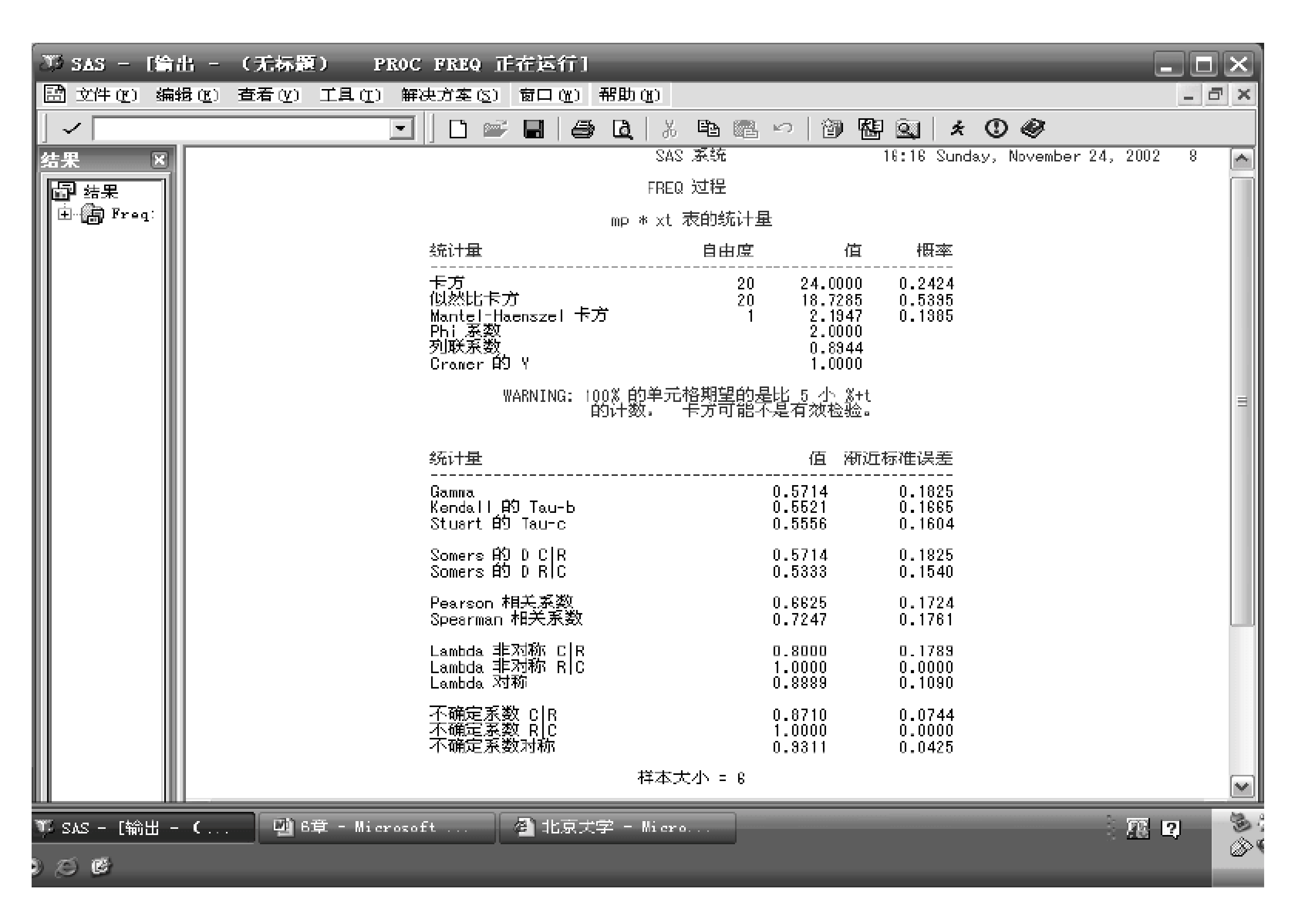


图 6.4 mp * xt 的结合程度

Pearson 系数,该系数 0.6625 比较大,说明两个样本相关。

2. 总体推论

但由于样本量小,无法进行总体推论。

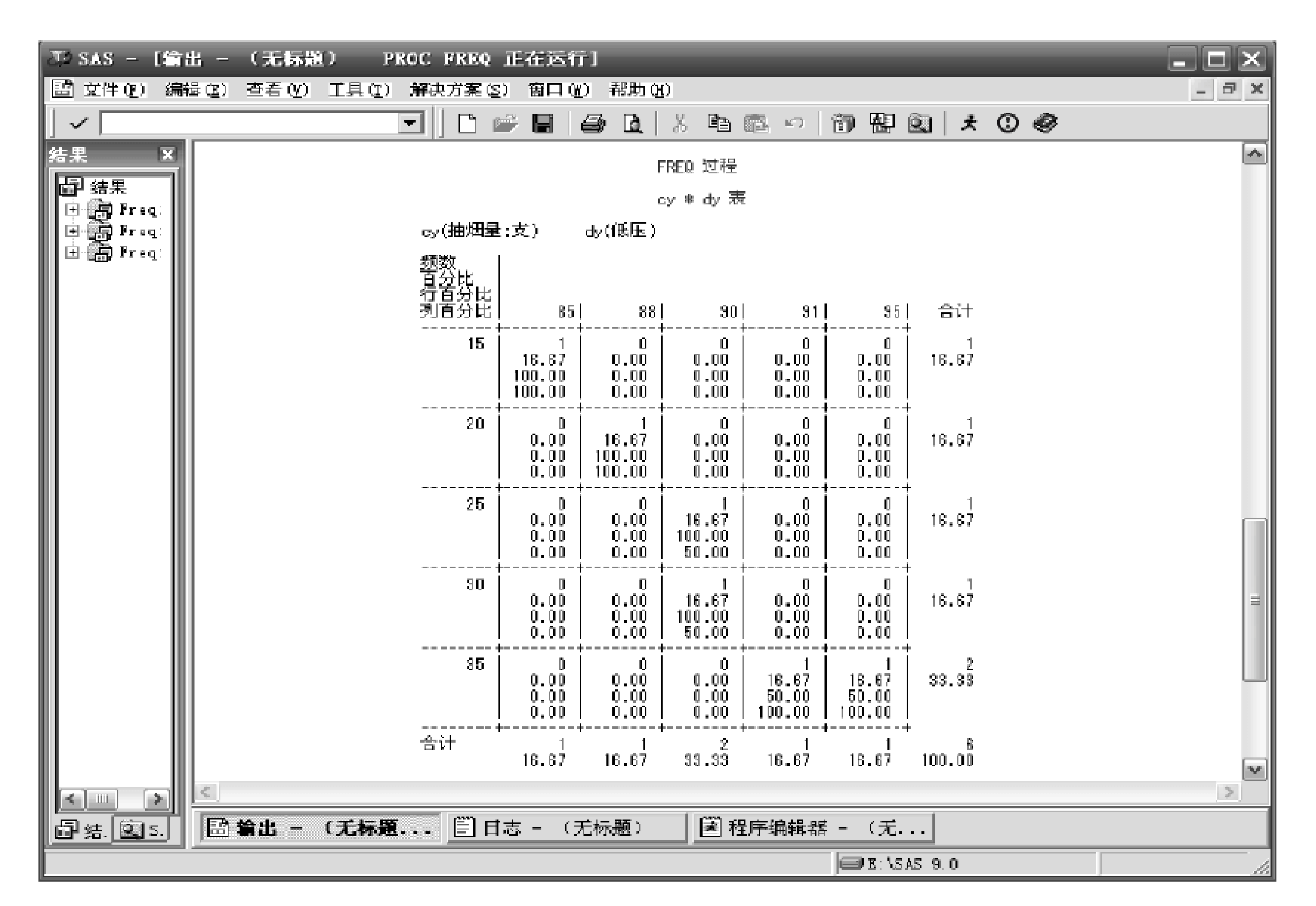
6.3.4 “定序-定序”双变量交叉汇总与结合测量

例 7:用程序 6.3 中的数据,命令语句见程序 6.4。

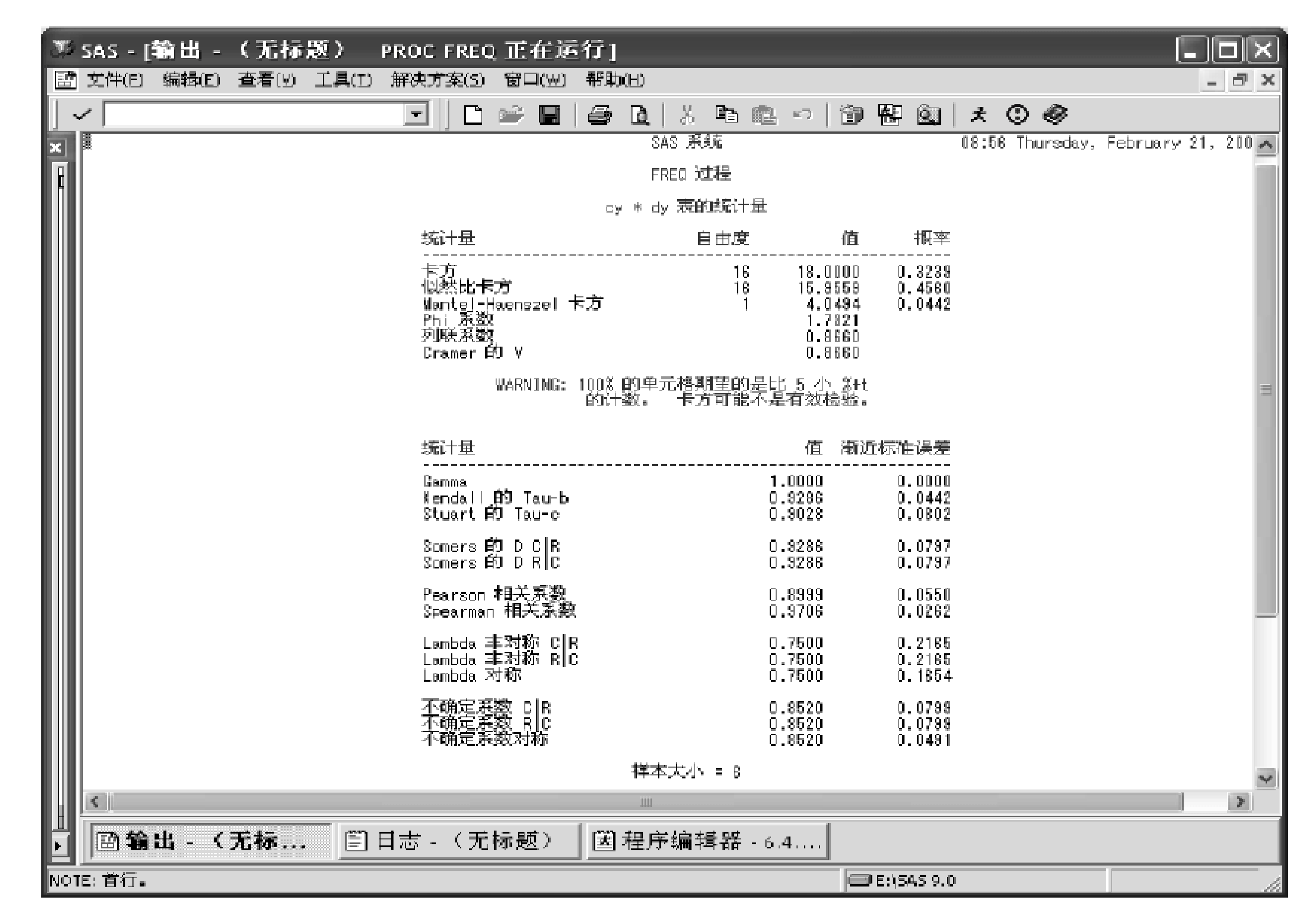
程序 6.4:“定序-定序”变量的测量。

```
DATA xt;
INPUT sex location cy dy mp xt 3.1;
LABEL location='地区' sex='性别' cy='抽烟量:支'
 mp='脉搏' dy='低压' xt='血糖';
CARDS;
1 1 15 85 66 8.5
2 1 20 88 68 7.3
1 2 30 90 70 8.6
2 2 25 90 70 7.8
1 1 35 95 75 9.2
2 2 35 91 76 9.0
;
PROC FORMAT;
  VALUE cyF LOW-10=1 11-20=2 21-HIGH=3;
  VALUE dyF LOW-85=1 86-90=2 91-HIGH=3;
  FORMAT cy cyF. Dy dyF.;
PROC FREQ; /* 产生图 6.5 * /
TABLE cy * dy/NOPRINT ALL;
```

运行程序 6.4 产生图 6.5 所示的结果。



(a) 交义汇总的单元表格



(b) 卡方检验

图 6.5 "定序-定序"变量的测量

结果分析:

(1) 单元内容分析

见图 6.5(a),可参考图 6.2 的分析。

(2) 双变量结合程度

从图 6.5(b)看,由于 cy * dy 属于"定序-定序"类型的数据,应该观察 Tau-b 系数,该

系数 0.9286 很大(接近 100%),说明两个样本强相关。

(3) 总体推论

H0:行列变量互为独立。

双变量结合测量的核心是卡方检验。它检验"行列变量互为独立"的总体推论。

检验:从图 6.5(b)看,样本小,而且卡方 18.0000,自由度 16,计算后的显著性水平为 0.3239,0.3239 大于 α 值 0.05,所以不能拒绝 H0,即总体上说,sex 和 edu 双变量互为独立。

6.4 再用 UNIVARIATE 过程详细描述单变量

UNIVARIATE 过程是对数字型变量的最详尽描述。它除了具有 Means、Sammary、Tabulate 等过程的功能及统计量外,还产生众数、中位数、峰度、偏度、四分位数、频率表等最详尽的统计量。

此外 UNIVARIATE 还产生以下统计量:

- 变量的极值;
- 几幅分布图,如茎叶图、盒图、正态概率图;
- 数据分布的正态性检验等。

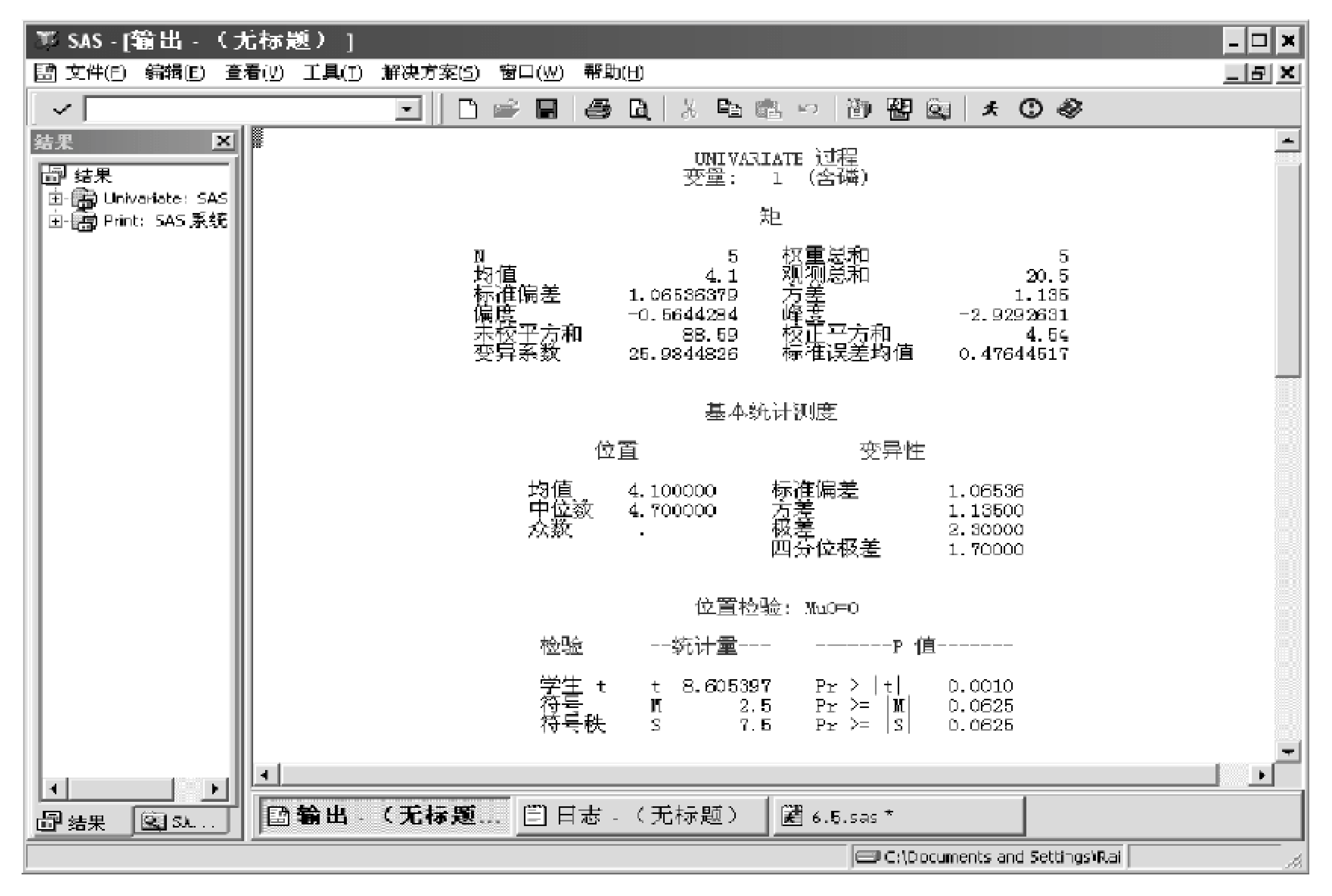
6.4.1 举例

程序 6.5:

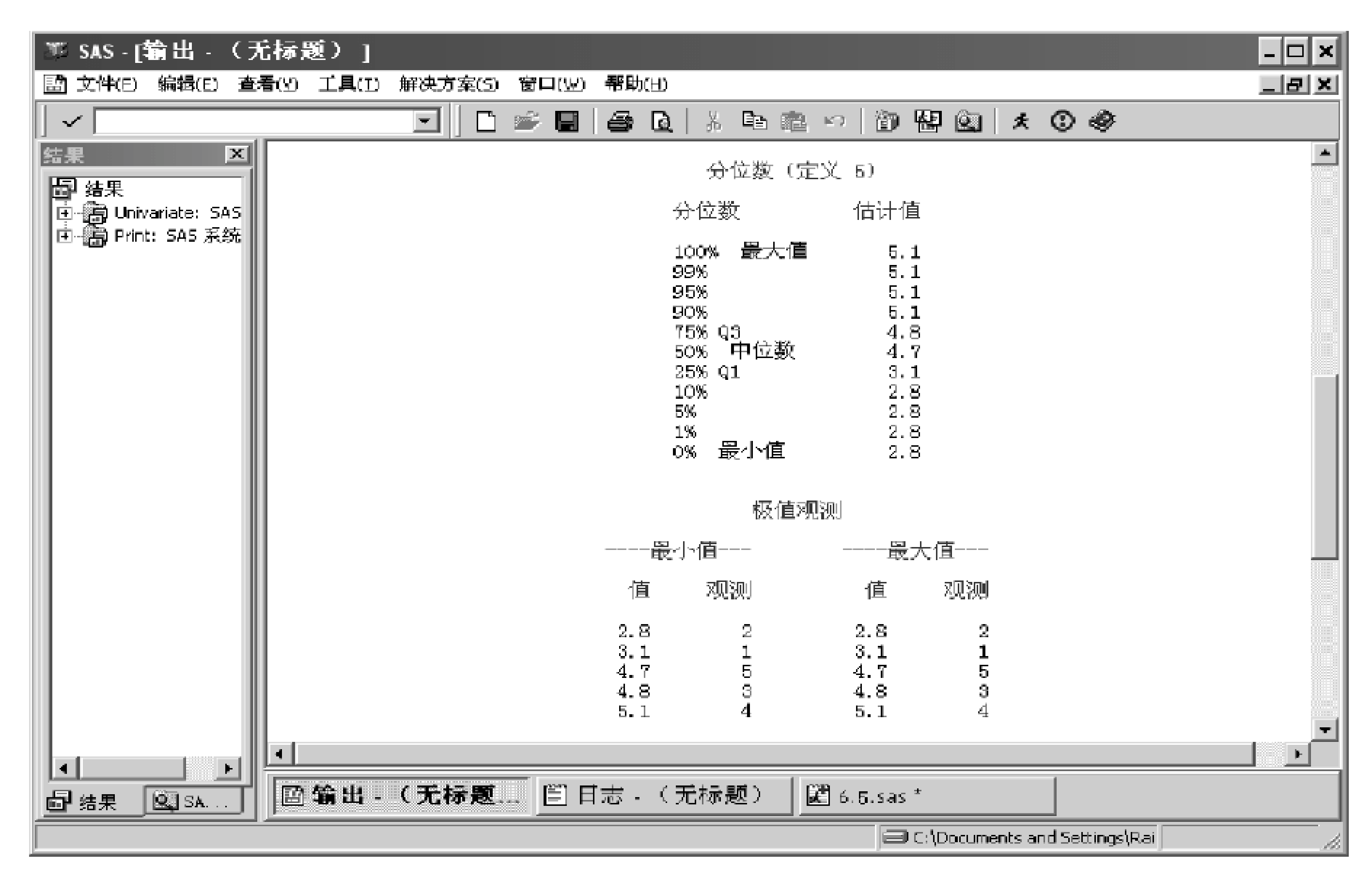
```
LIBNAME LB 'D:\';
DATA LB.xt2;
LABEL edu='文化水平' sex='性别' l='含磷' xt='血糖';
INPUT id sex edu xt l ;
CARDS;
001 1 1 8.1 3.1
002 2 2 9.1 2.8
003 1 3 9.0 4.8
004 2 3 8.7 5.1
005 1 2 . 4.7
006 2 . 6.2 .
PROC UNIVARIATE DATA=LB.xt2 NORMAL;
  VAR L xt ; /* 血糖和磷的描述统计 * /
  OUTPUT OUT=UNI1;
PROC PRINT;
RUN;
```

运行程序 6.5 可产生图 6.6 所示的结果。

从图 6.6(a)看,这些被访者的血液中平均含磷 4.1,大于正常值。标准偏差为 1.06536379 表明变异性突出,且为偏离均值之上。

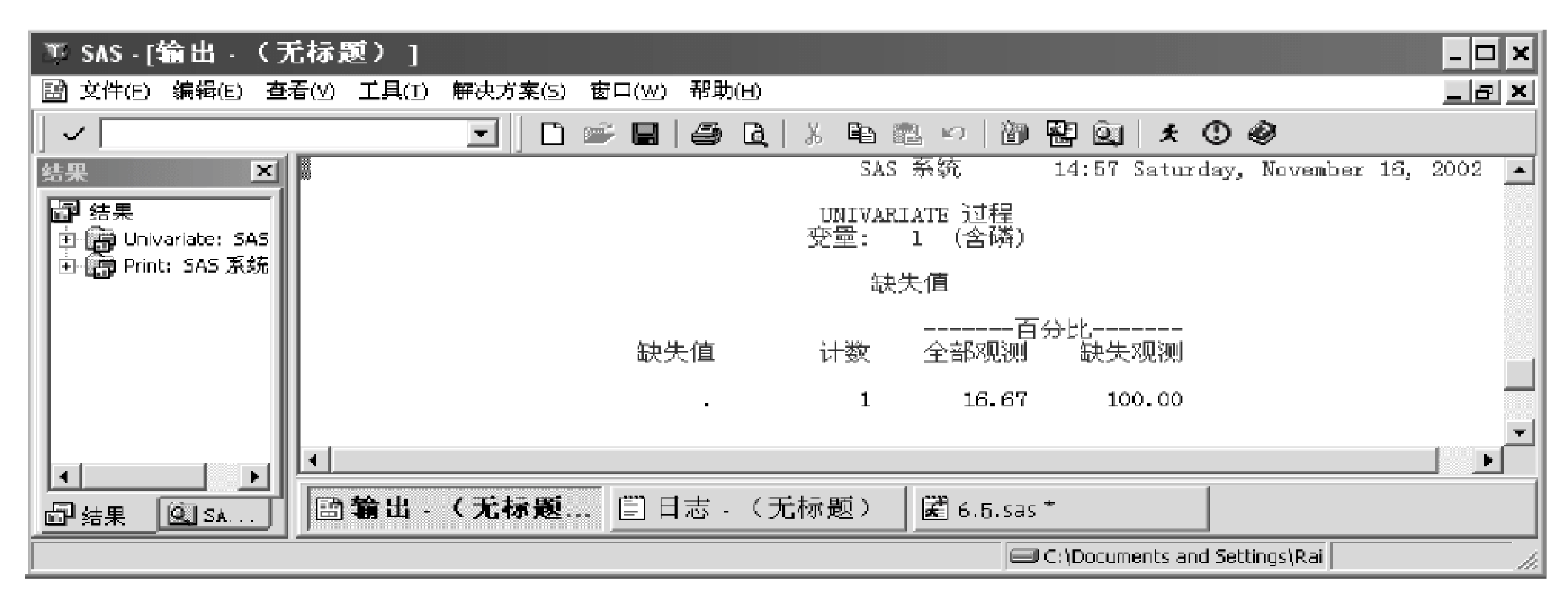


(a) 均值等基本描述



(b) 四分位数和最大最小值

图 6.6 含磷数据的描述



(c) 缺失值统计



(d) 正态性检验

图 6.6（续）

而且学生 t 分布值为 t=8.605397，检验概率“Pr ＞|t|”值为 0.0010，很显著。

图 6.6(b)显示四分位数和最大最小值各 5 个，这对于分析数据的异常值很有帮助，一般要删除数据中的这些最大、最小值。

图 6.6(c)显示出有 1 个人的数据缺失，缺失值占总个案的 16.67%，比例很大应该删除有缺失值的个案。

从图 6.6(d)正态性检验看，Shapiro-Wilk 的 W 值 0.828117 太小，其检验的显著性水平 0.1347 又太大，即大于 α 值 0.05，所以不能拒绝 H0 假设，表明数据基本上呈现正态分布。

说明：按此法可对其他数字型变量进行描述。

6.4.2 UNIVARIATE 过程命令

1. 过程命令的格式

```
PROC UNIVARIATE DATA=已建的数据集 a PLOT FREQ MORMAL;
    VAR v1 v2;
[BY v3;]
```

```
[FREQ v4;]
[ID v5;]
OUTPUT OUT=输出数据集 V=v11 v12 PCTLPTS=25 50 PCTLNAME=P25 P50 ;
```

2. 选项说明

- DATA＝已建的数据集 a：数据集的名称（如 a），与 DATA 步指定的名称要一致。
- PLOT：只要求画出散点图。
- FREQ：只要求输出频率表。
- MORMAL：只要求输出正态性检验的统计量。
- BY v3：要求按 v3 变量值排序个案，可省略。
- FREQ v4：要求计算 v4 变量的频数分布，可省略。
- ID v5：用 v5 变量值的前 8 个字符命名 5 个最小值和 5 个最大值，不可省略。
- V＝v11 v12：v11 变量名与“VAR v1 v2;”中的 v1 对应，v12 变量名与“VAR v1 v2;”中的 v2 对应。
- P25 P50：P25 变量名与“PCTLPTS＝25 50;”中的 25 对应，表示百分位数的名称。P50 变量名与“PCTLPTS＝25 50;”中的 50 对应，表示百分位数的名称。

例 8：百分位数的名称的输出。

程序 6.6：将上例的 OUTPUT 语句改为“OUTPUT OUT＝输出数据集 V＝v11 v12 PCTLPTS＝25 50”。

```
DATA xt2;
LABEL edu='文化水平' sex='性别' l='含磷' xt='血糖';
INPUT id sex edu xt l @@ ;
CARDS;
001 1 1 8.1 3.1 002 2 2 9.1 2.8
003 1 3 9.0 4.8 004 2 3 8.7 5.1
005 1 2 . 4.7 006 2 . 6.2 .
PROC UNIVARIATE DATA=xt2 NORMAL;
VAR xt; /* 血糖描述统计 */
ID id; /* 用 id个案号的前 8个字符命名 5个最小值和 5个最大值。便于查找 */
OUTPUT OUT=OU2 N=L1 PCTLPRE=STD PCTLPTS=25 50
    PCTLNAME=P25 P50 ;
PROC PRINT;
RUN;
```

运行程序 6.6 可产生图 6.7 所示的结果。

从图 6.7 看，用“ID id;”语句时可以一目了然地查出哪些个案是最大值和最小值，而且，用“PCTLPRE＝STD PCTLPTS＝25 50”语句可以观察到第几个百分位数的值。

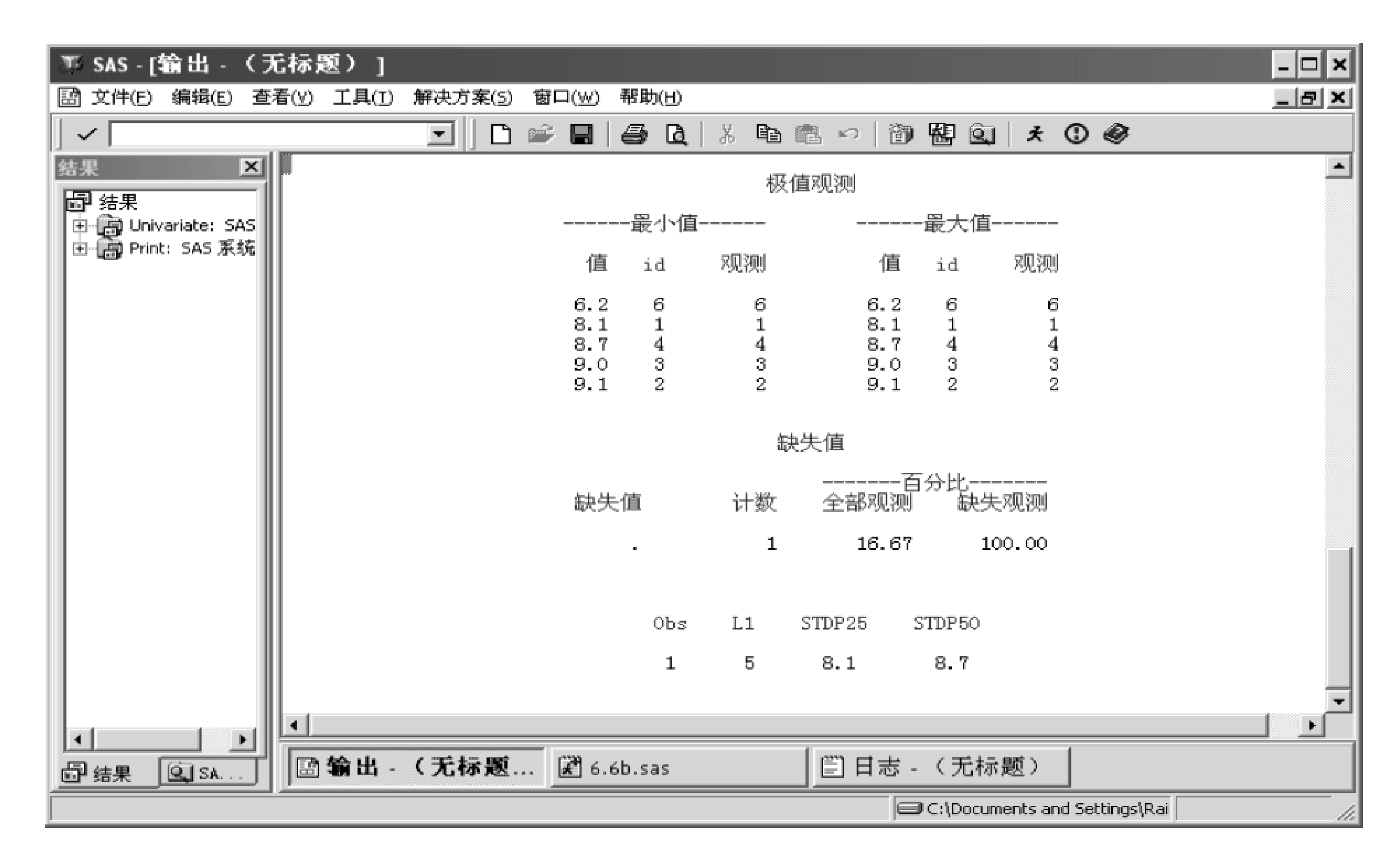


图 6.7 "ID id;"及"PCTLPRE=STD PCTLPTS=25 50"语句的输出

6.4.3 计算方法

1. 统计检验

T 检验：

H0：总体均值为 0。

检验值：采用 Student 的 $T=(x^{-}N^{1/2})/S$

由 T 值通过查表或计算获得显著性水平 P 值，P 值<α 值 0.05，则拒绝"总体均值为 0"的 H0 假设。

2. 正态性检验

当指定"PROC UNIVARIATE NORMAL;"语句时，UNIVARIATE 过程假设该样本数据是取自正态分布的总体，并且给出一个检验的统计量。

当样本量小于等于 2000 时，则进行 Shapiro-Wilk 检验的统计量 W；当样本量大于 2000 时，则采用 Kolmogorov 检验的统计量 D。

当样本量大于 6 时，Shapiro-Wilk 检验的统计量 W 的显著性水平，由 Royston 近似正态变换获得。

6.5 进一步用 PROC CHART 过程描述单变量

PROC CHART 过程除了产生频率表外，还产生水平直方图(或水平条形图)、垂直直方图(或垂直条形图)、圆形图、立体图(block plot)和星形图。

直方图比条形图更能充分地描述数据。条形图只能描述性别等 nominal(标称、定类)数据,它测量不到区间。直方图适宜于 interval(区间、定距)数据和 ratio(比例、定比)数据。

6.5.1 PROC CHART过程命令

过程命令的格式如下:

```
PROC CHART DATA=已建的数据集; [BY v1;]
HBAR    v2 [选项 1];
VBAR    v2 [选项 1];
PIE     v2 [选项 1];
BLOCK   v2 [选项 1];
STAR    v2 [选项 1];
```

6.5.2 CHART的选项 1

1. HBAR~STAR 中的公共选项 1

选项 1 如下:

```
HBAR v2 /DISCRETE;          /* 产生标称(定类)变量的离散值 */
        /MISSING;           /* 把缺失值当作有效值计算 */
        /SUMVAR=v2;         /* 计算变量 v2 的均值、总和、频数 */
        /MIDPOINTS=1 2 3;   /* 中心值,可以是 1 2 3 4 5 或 2 4 6 等 */
        /FREQ v;            /* 按变量 v 分类。将同类中变量 v 的和当作频数 */
        /AXIS=最小值 最大值;   /* 表示 FREQ、PCT、CFREQ、CPCT、SUM、MEAN 轴上的最小值和最大值。
                                若只指定一个值,则被当作最大值 */
        /TYPE=FREQ;         /* 默认为用每个条形或区段,表示某变量值或范围所出现的频数 */
        /TYPE=SUM;          /* 与选项 SUMVAR=v2 连用时,则按 v2 分类计算出 v2 值之和 */
        /TYPE=MEAN;         /* 与选项 SUMVAR=v2 连用时,则按 v2 分类计算 v2 值的均值 */
```

例 9:在程序 6.3 后面增加 HBAR xt/SUMVAR=xt 命令,见程序 6.7。

程序 6.7:

```
DATA xt;
LABEL location='地区' sex='性别' cy='抽烟量:支' mp='脉搏'
    dy='低压' xt='血糖';
INPUT sex location cy dy mp xt;
CARDS;
1 1 15 85 66 8.5
2 1 20 88 68 7.3
1 2 30 90 70 8.6
2 2 25 90 70 7.8
1 1 35 95 75 9.2
2 2 35 91 76 9.0
;
```

```
PROC CHART; /* 产生图 6.8* /
  HBAR xt/SUMVAR=xt TYPE=MEAN;
PROC CHART; /* 产生图 6.9* /
  HBAR xt/GROUP=sex SUMVAR=xt TYPE=MEAN;
PROC CHART; /* 产生图 6.10* /
  HBAR xt/GROUP=sex SUBGROUP=location SUMVAR=xt TYPE=MEAN;
```

运行程序 6.7 产生图 6.8 至图 6.10 所示的结果(平均血糖的直方图)。

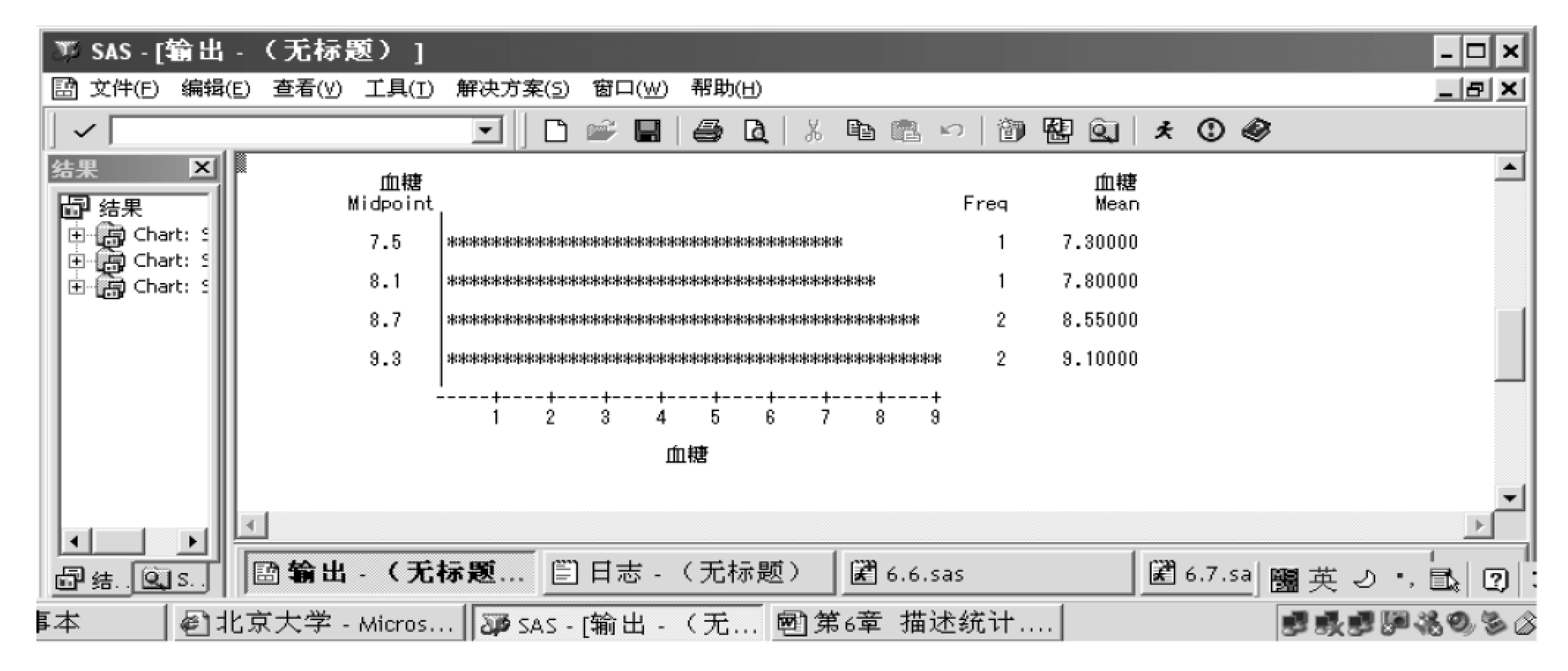


图 6.8 血糖综合直方图

图 6.8 是由“PROC CHART; HBAR xt/SUMVAR=xt TYPE=MEAN;”过程命令产生的综合直方图。有 1 人血糖平均 7.30000,有 1 人血糖平均 7.80000,有 2 人血糖平均 8.55000,有 2 人血糖平均 9.10000,都是高血糖。

图 6.9 是由“PROC CHART; HBAR xt/ GROUP=sex SUMVAR=xt TYPE=MEAN;”过程命令产生的综合直方图。

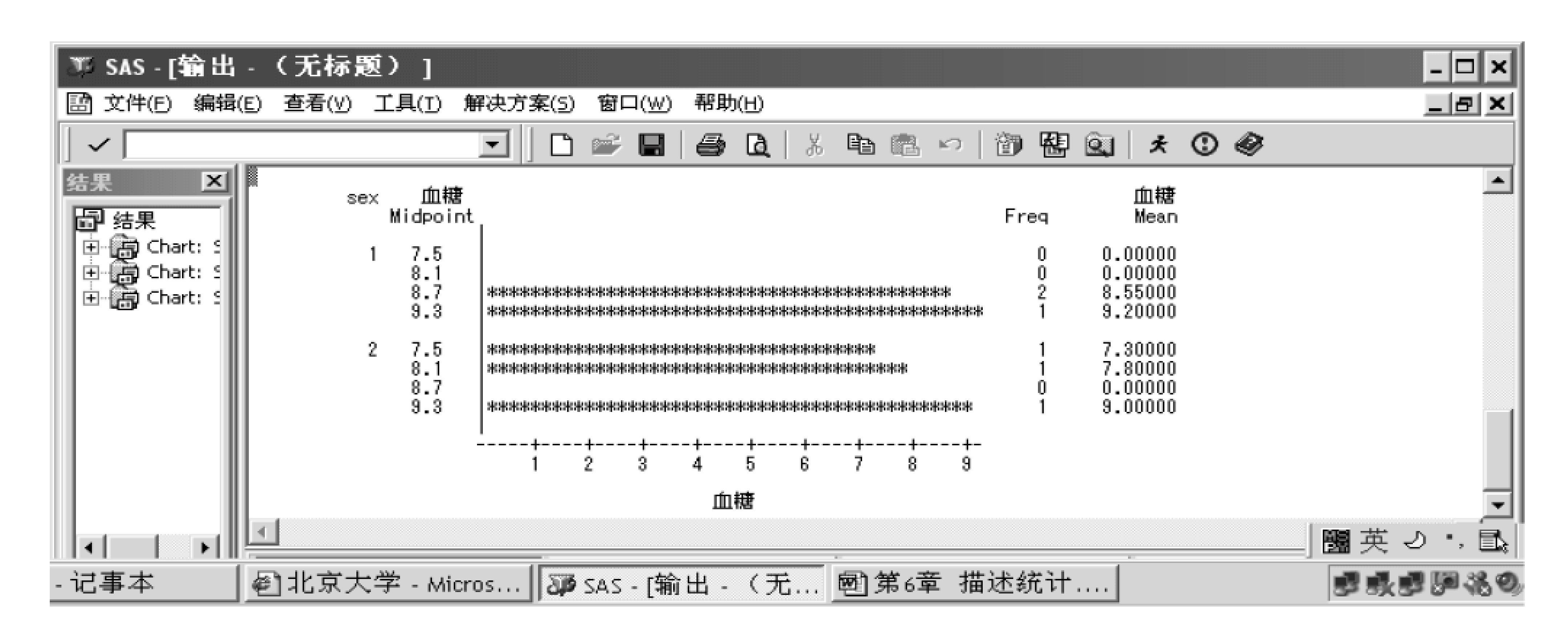


图 6.9 按地区分的血糖直方图

从图 6.9 看:先按性别粗分为 2 组。区间中点(Midpoint)的 8.7 表示血糖 8.4~9.0 的有 2 人,区间的中点 9.3 表示血糖 9.0~9.5 的有 1 人。而不是血糖 8.7 的有 2 人,血

糖 9.3 的有 1 人。

余者以此类推。

图 6.10 是由以下命令产生的：

```
PROC CHART; HBAR xt/GROUP= sex SUBGROUP= location SUMVAR= xt TYPE= MEAN;
```

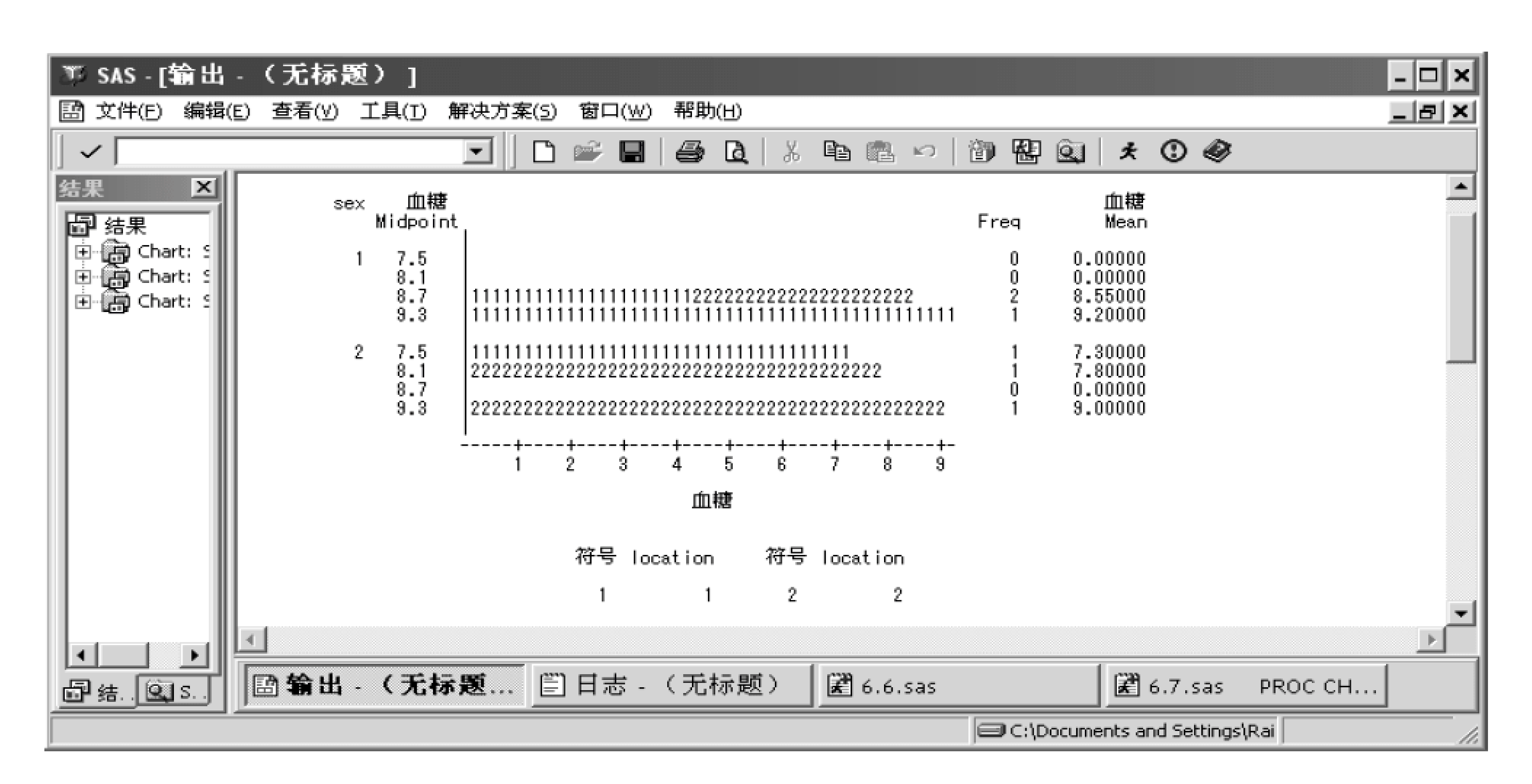


图 6.10 按性别分的并进一步按地区分的血糖直方图

先按性别粗分为 2 组，每组的条形再详细按 location＝1 和 location＝2 两种水平画出条形。所以最后这种直方图比较直观。

2. 专用于 HBAR、VBAR 和 BLOCK 图形中的选项

```
PROC CHART;
  HBAR xt/GROUP= sex
          SUBGROUP= location
          LEVEL= n; /* 当"HBAR xt"中的变量 xt 是连续型的变量时可用 LEVEL= n
                    指定要输出几条条形 */
```

例 10：指定 LEVEL＝n，见程序 6.8。

程序 6.8：

```
DATA xt;
LABEL location= '地区' sex= '性别' cy= '抽烟量:支' mp= '脉搏'
      dy= '低压' xt= '血糖';
INPUT sex location cy dy mp xt;
CARDS; /* 血糖 5单位以下,磷 3单位以下为正常 */
1 1 15 85 66 8.5
2 1 20 88 68 7.3
1 2 30 90 70 8.6
2 2 25 90 70 7.8
1 1 35 95 75 9.2
```

```
2 2 35 91 76 9.0
;
PROC FORMAT;
PROC CHART;
  BLOCK xt/SUBGROUP= sex SUMVAR= xt TYPE= MEAN ;/* 产生图 6.11 上半图 */
  BLOCK xt/SUBGROUP= sex LEVELS= 2 SUMVAR= xt TYPE= MEAN; /* 产生图 6.11 下半图 */
```

运行程序 6.8 产生图 6.11 所示的结果。

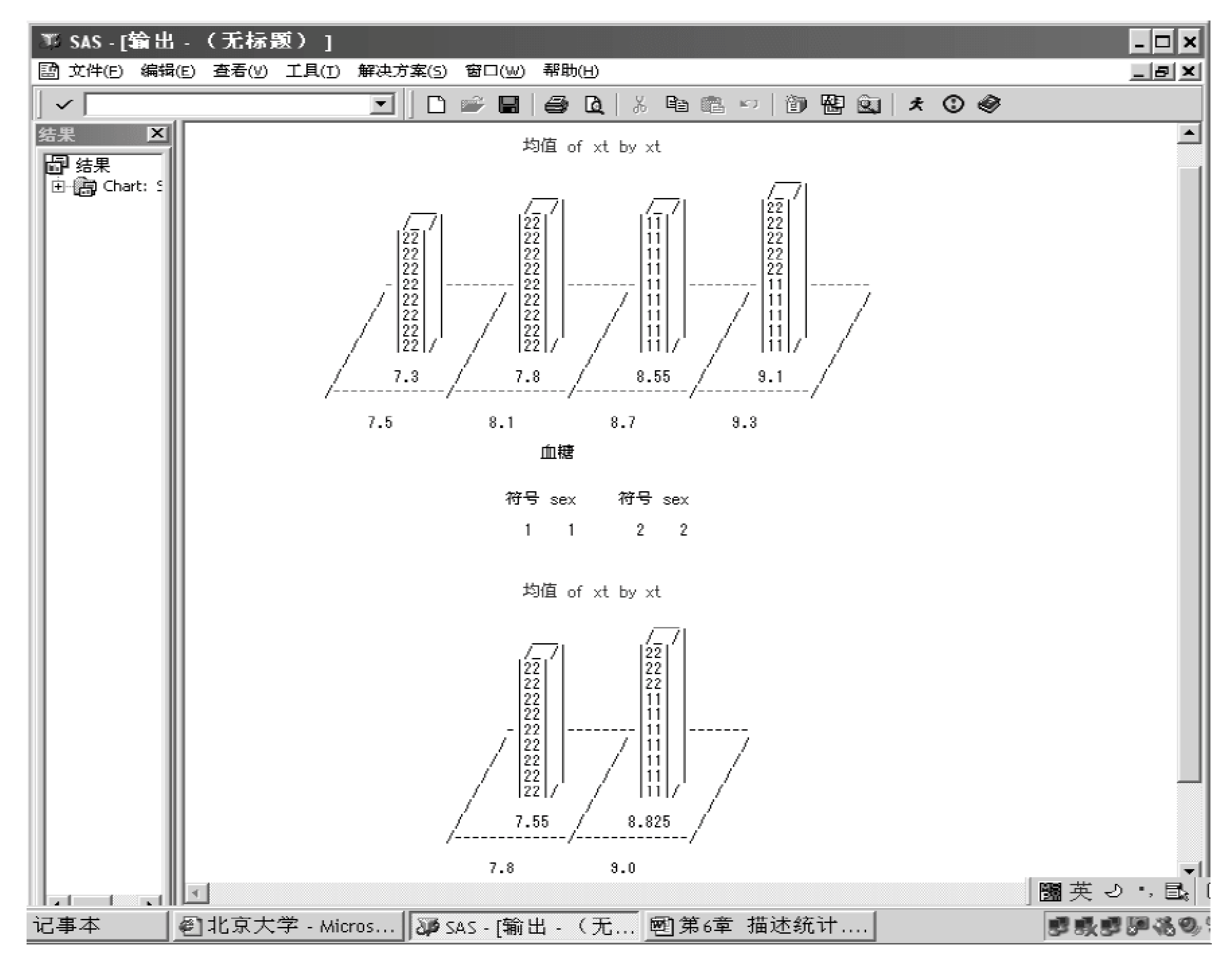


图 6.11 有无 LEVELS=选项的立体图

从图 6.11 看到,上半图没有指定 LEVELS=2,所以按默认的刻度画出 4 条条形。因为图 6.11 的下半图指定了 LEVELS=2,所以只画出 2 个条形。男性(用 1 表示)比女性(用 2 表示)平均血糖高。不过还是不清晰,可再观察水平直方图(图 6.12)。

3. 专用于 HBAR、VBAR 图形中的选项

```
PROC CHART;
  HBAR xt/GROUP= sex
        SUBGROUP= location
        LEVEL= n    /* 当“HBAR xt”中的变量 xt 是连续型的变量时可用 LEVEL= n 指定要输出几条条
                       形 */
        REF= n;     /* 产生一条参考线,它与 TYPE= 选项连用,当 TYPE= FREQ 时它表示频数,当 TYPE
                       = PCT 时表示百分比,当 TYPE= SUM|MEAN 时,分别表示和数或均值 */
```

例 11:指定 LEVEL=n,REF=2,见程序 6.9。

程序 6.9：

```
DATA xt;
LABEL location='地区' sex='性别' cy='抽烟量:支/日'
    mp='脉搏' dy='低压' xt='血糖';
INPUT sex location cy dy mp xt;
CARDS; /* 磷 3单位以下为正常 */
1 1 15 85 66 8.5
2 1 20 88 68 7.3
1 2 30 90 70 8.6
2 2 25 90 70 7.8
1 1 35 95 75 9.2
2 2 35 91 76 9.0
;
PROC CHART;
  HBAR sex/SUMVAR=xt LEVEL=2;
```

运行程序 6.9 产生图 6.12 所示的结果。

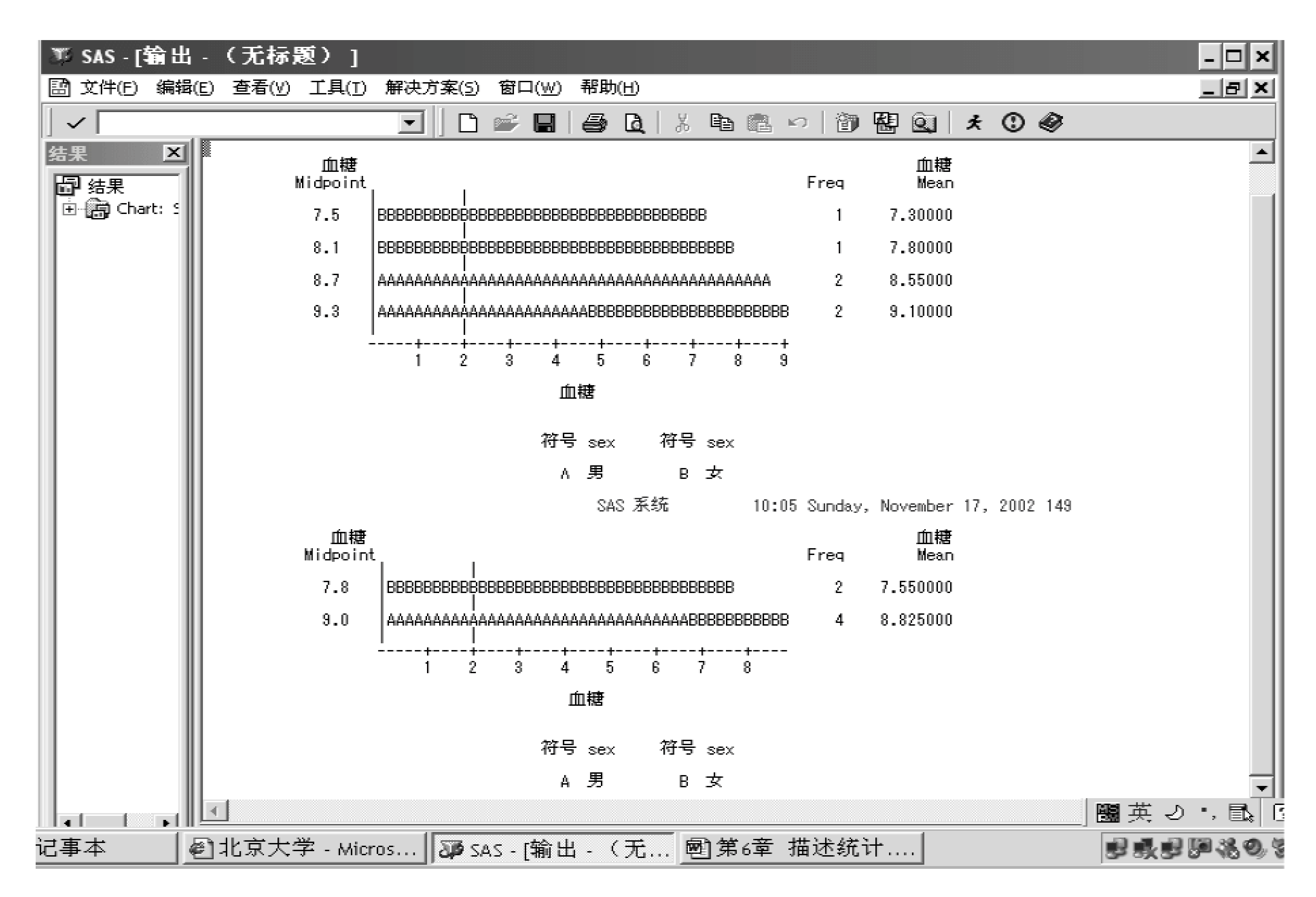


图 6.12　平均血糖的直方图

比较图 6.12 上半图和下半图可以看出：平均血糖虽然没有实质性的改变，但图 6.12 的上半图没有指定 LEVELS=2，所以按默认的刻度画出。又因为图 6.12 的下半图指定了 LEVELS=2，所以只画出 2 个条形。男性（用 A 表示）比女性（用 B 表示）平均血糖高。

参考线=2 在这里显得不太重要。

4. 专用于 HBAR 图形中的选项

```
PROC CHART;
HBAR /TYPE=FREQ;       /* 在水平直方图的右侧显示每个条形的频数 */
    /TYPE=CFREQ;       /* 在水平直方图的右侧显示每个条形的累积频数 */
    /TYPE=SUM;         /* 见 6.5.2 节,与选项 SUMVAR=v2 连用时,按 v2 分类计算 v2 的和 */
    /TYPE=MEAN;        /* 见 6.5.2 节,与选项 SUMVAR=v2 连用时,按 v2 分类计算 v2 的均值 */
    /TYPE=PERCENT;     /* 见 6.5.2 节,计算个案的百分比 */
    /TYPE=CPERCENT;    /* 见 6.5.2 节,计算个案的累积百分比 */
    /TYPE=NOSTAT;      /* 不显示统计量 */
```

说明:

- 有/TYPE=但无/SUMVAR=选项时,CHART 过程输出 FREQ、CFREQ、PERCENT 和 CPERCENT 值。
- 有/TYPE=且有/SUMVAR=选项时,CHART 过程输出 FREQ、MEAN 值。
- 有/TYPE=SUM 时 CHART 过程输出 FREQ 和 SUM 值。

例子见程序 6.10。

程序 6.10:

```
DATA xt;
LABEL location='地区' sex='性别' cy='抽烟量:支/日'
  mp='脉搏' dy='低压' xt='血糖';
INPUT sex location cy dy mp xt;
CARDS;
1 1 15 85 66 8.5
2 1 20 88 68 7.3
1 2 30 90 70 8.6
2 2 25 90 70 7.8
1 1 35 95 75 9.2
2 2 35 91 76 9.0
;
PROC CHART;
  VBAR xt/GROUP=sex TYPE=MEAN;
  HBAR xt/GROUP=sex TYPE=MEAN LEVELS=2;
```

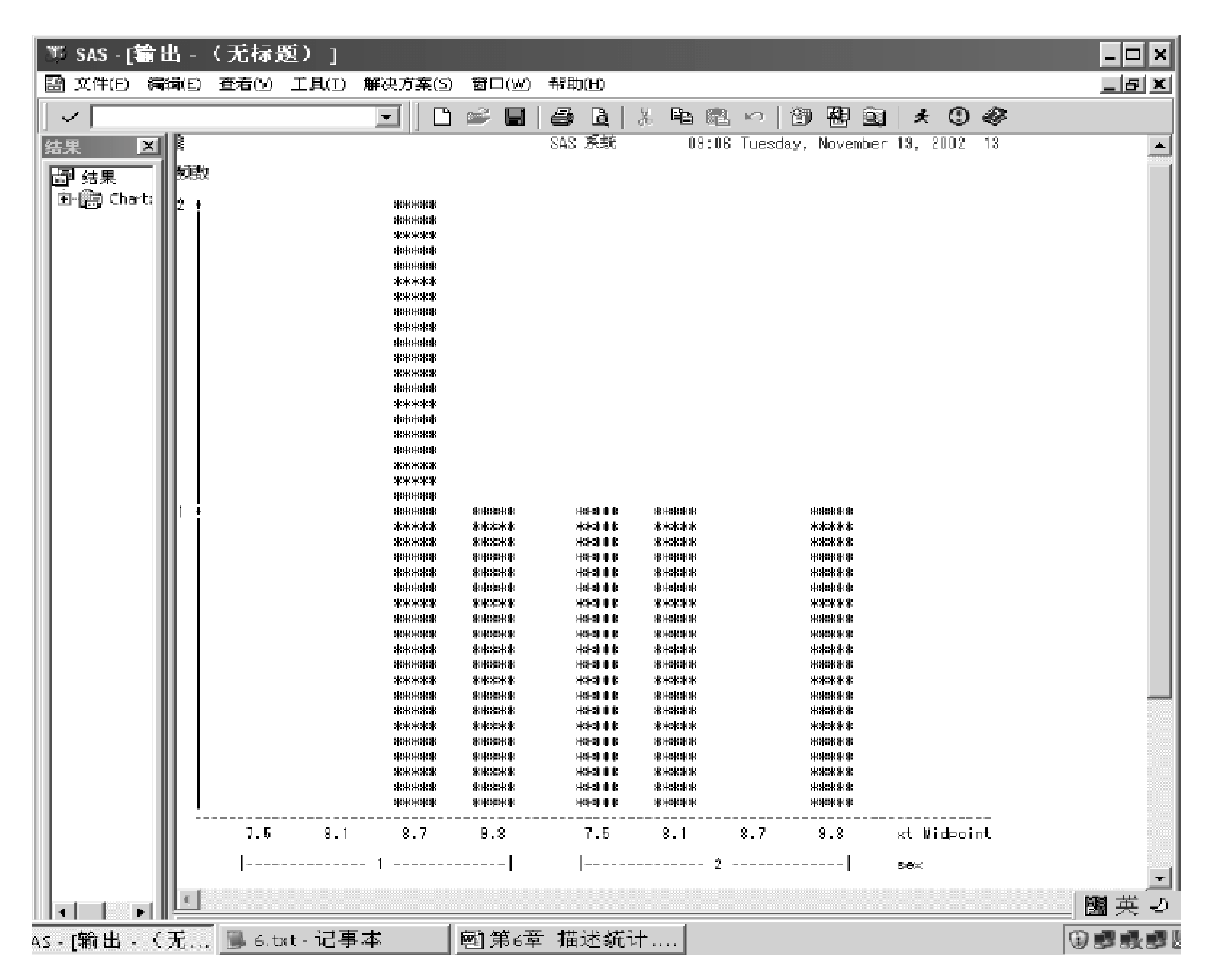
运行程序 6.10 产生图 6.13 所示的结果。

从图 6.13(b)看,有/TYPE=但无/SUMVAR=选项时,CHART 过程输出 FREQ、CFREQ、PERCENT 和 CPERCENT 值,但不产生 MEAN 值。

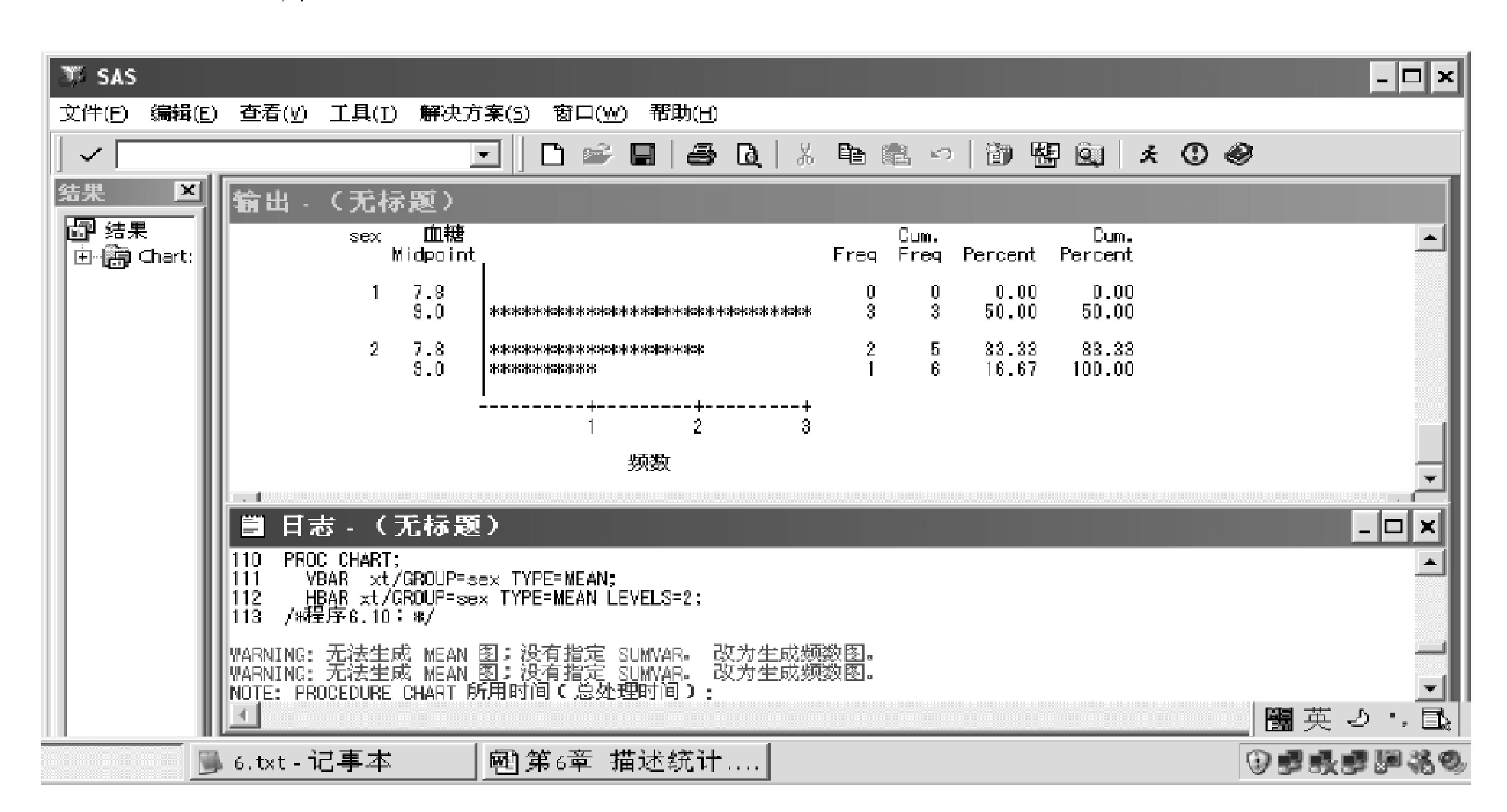
5. 绘制圆形图

程序 6.11:

```
DATA xt;
LABEL location='地区' sex='性别' cy='抽烟量:支/日'
```



(a) 由 “ VBAR xt/GROUP=sex TYPE=MEAN;” 产生的垂直直方图



(b) 由 “ HBAR xt/GROUP=sex TYPE=MEAN LEVELS=2;” 产生的水平直方图

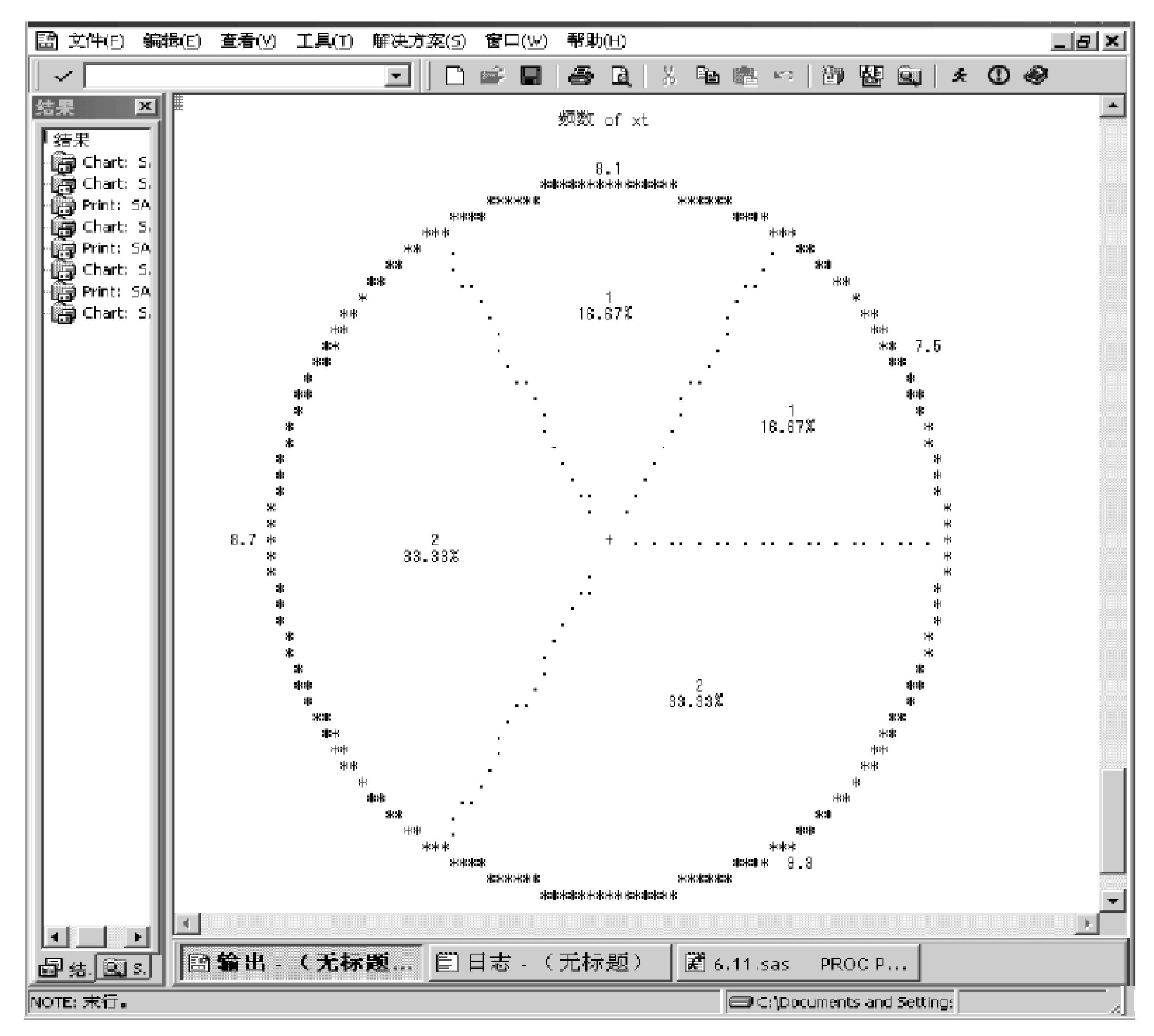
图 6.13　两种直方图

```
  mp='脉搏' dy='低压' xt='血糖';
INPUT sex location cy dy mp xt;
CARDS;
1 1 15 85 66 8.5
2 1 20 88 68 7.3
1 2 30 90 70 8.6
```

```
2 2 25 90 70 7.8
1 1 35 95 75 9.2
2 2 35 91 76 9.0
;
PROC CHART;
    PIE xt/MISSING SUMVAR= xt TYPE= FREQ TYPE= MEAN;
PROC PRINT;
PROC CHART;
    PIE xt/MISSING TYPE= FREQ;
PROC PRINT;
```

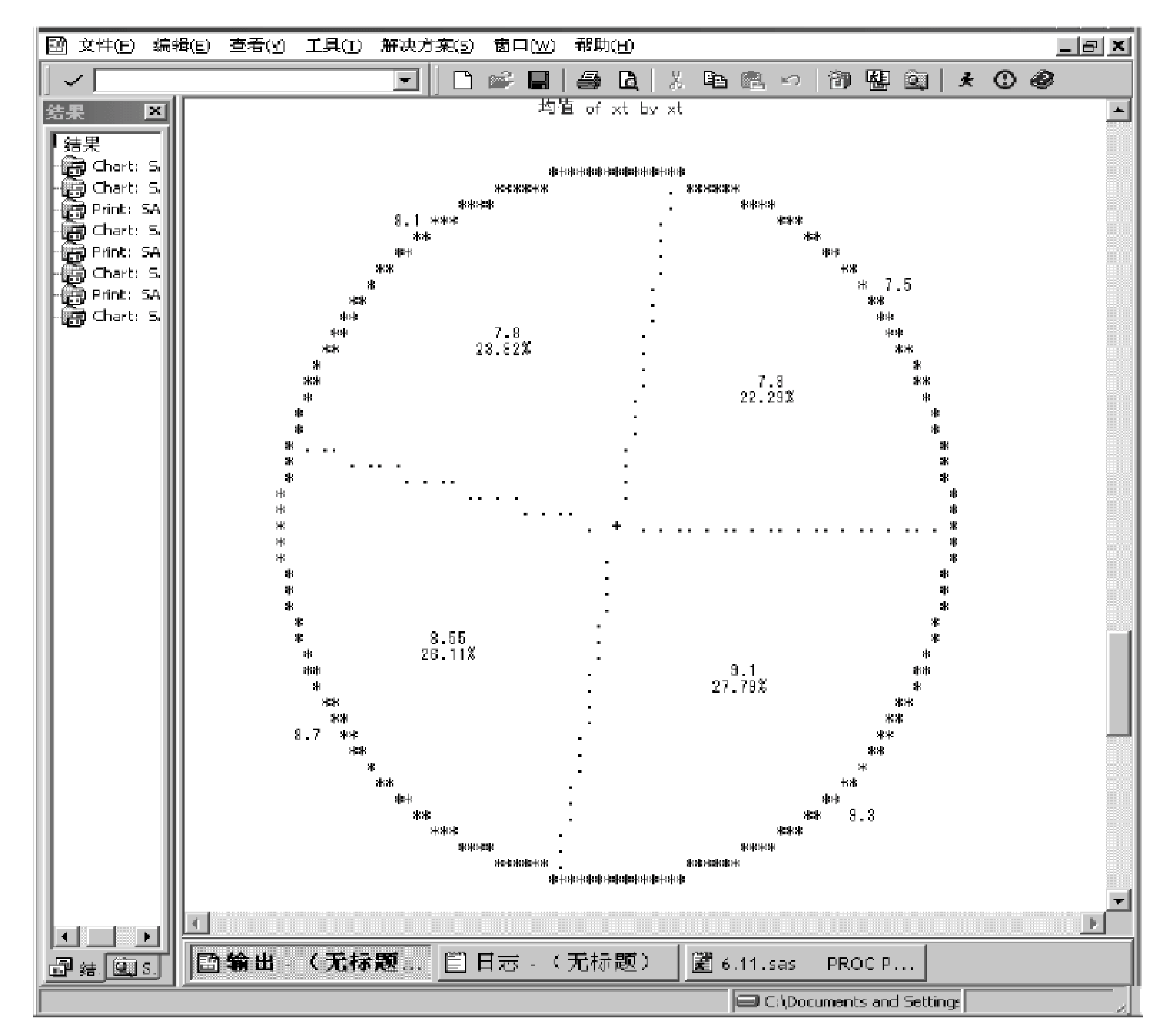
运行程序 6.11 产生图 6.14 所示的结果。

从图 6.14 看,难于既产生圆形图中的频数又显示出均值,只能分开产生输出。平均血糖为 7.5 者有 1 人,占总人数 16.67%。



(a) 由"PIE xt/MISSING TYPE=FREQ;"命令产生的圆形图

图 6.14　两种圆形图



(b) 由“ PIE xt/MISSING SUMVAR=xt TYPE=FREQ TYPE=MEAN;” 命令产生的圆形图

图 6.14（续）

6.6 用 MEANS 过程比较两个均值

MEANS 过程是对数字型变量计算各组(子总体)的均值,但对字符型变量无效。

6.6.1 应用实例

例 12:计算男女两组血糖平均含量,见程序 6.12。

程序 6.12:

```
DATA xt3;
LABEL location='地区' sex='性别' cy='抽烟量:支/日'
    mp='脉搏' dy='低压' xt='血糖';
INPUT sex location cy dy mp xt;
CARDS;
1 1 15 85 66 8.5
2 1 20 88 68 7.3
1 2 30 90 70 8.6
2 2 25 90 70 7.8
1 1 35 95 75 9.2
2 2 35 91 76 9.0
```

```
;
PROC SORT;BY sex;
PROC MEANS;
     BY sex;
   VAR xt;
   OUTPUT OUT=STA1;
```

运行程序 6.12 产生图 6.15 所示的结果。

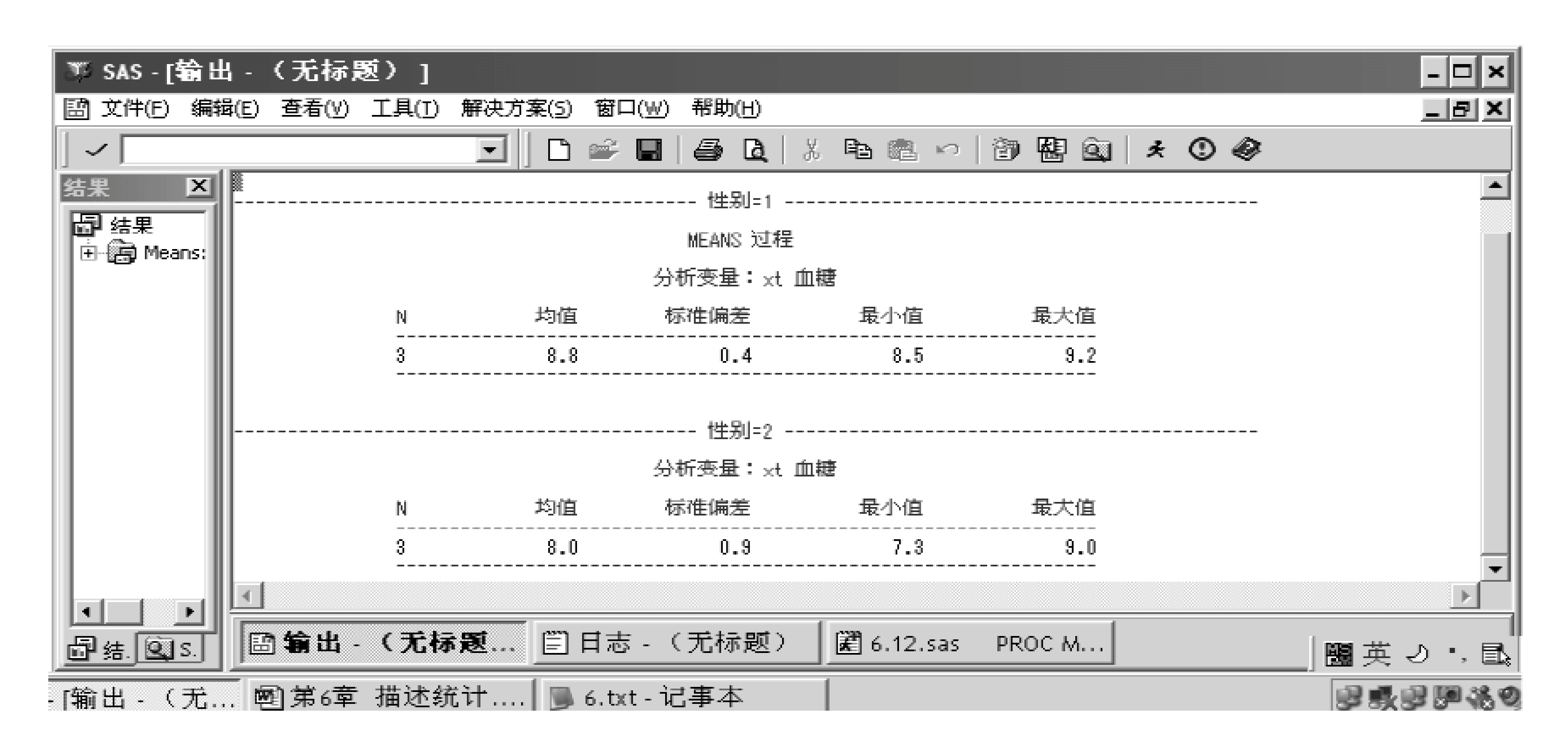


图 6.15 MEANS 过程默认格式的输出

从图 6.15 看，男性一组平均血糖 8.8，最高；标准偏差 0.4，相对较低。女性一组平均血糖 8.0，稍微低一些；标准偏差 0.9，相对较高。从理论上看，均值高、标准偏差又大的一组才是真正的高。

一般地说，用 MEANS 过程的默认格式，其输出足够描述数据。

6.6.2 MEANS 过程命令

1. MEANS 过程命令的格式

```
PROC MEANS DATA=已建的数据集 MAXDEC=0~8 VARDEF=N|DF|WEIGHT|WDF
  N=    MIN=    MAX=    MEAN=    STD=    STDERR=    SUM=    VAR=    USS=    CSS=
  NMISS  =   RANGE=
  T=    PRT=    SUMWGT=    CV=    SKEWNESS KURTOSIS=    ;
VAR v1 v2;
OUTPUT OUT=输出数据集如 STA1;
```

2. 格式说明

```
MAXDEC=0~8       /* 小数位。默认为小数位占 2 位 */
VARDEF=N         /* 用个案数作为除数 */
       DF        /* 用(自由度 N-1)作为除数。默认 */
       WEIGHT    /* 用“权重和”作为除数 */
       WDF       /* 用“权重和-1”作为除数 */
```

```
N=              /* 显示取多少个个案 * /
MIN=            /* 显示最小值 * /
MAX=            /* 显示最大值 * /
MEAN=           /* 显示均值 * /
STD=            /* 显示标准偏差 * /
STDERR=         /* 显示标准误差 * /
```

以上几个统计量为默认的,一般够用。

3. 举例

程序 6.13:计算男女两组血糖平均含量,并存储均值变量。

```
DATA xt3;
LABEL location='地区' sex='性别' cy='抽烟量:支/日'
    mp='脉搏' dy='低压' xt='血糖';
INPUT sex location cy dy mp xt;
CARDS;
1 1 15 85 66 8.5
2 1 20 88 68 7.3
1 2 30 90 70 8.6
2 2 25 90 70 7.8
1 1 35 95 75 9.2
2 2 35 91 76 9.0
;
PROC SORT;BY sex;
  PROC MEANS;
    BY sex;
   VAR xt dy;
    OUTPUT OUT=STA2 MEAN=m1 m2 STD=s1 s2;
```

运行程序 6.13 产生图 6.16 所示的结果。

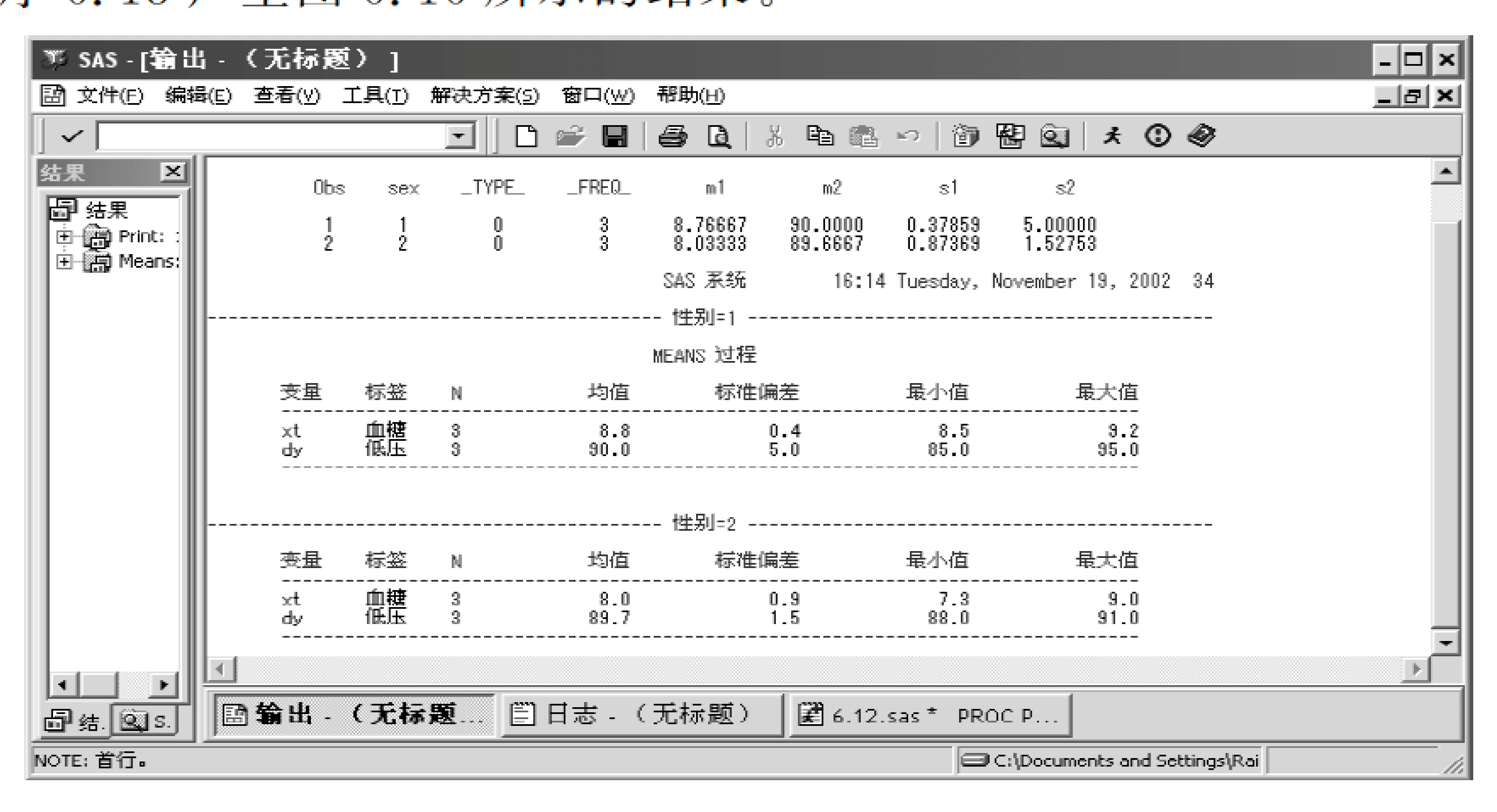


图 6.16 由“OUTPUT OUT=STA2 MEAN=m1 m2 STD=s1 s2;”产生的结果

分析：

MEAN＝m1 对应 VAR xt，STD＝s1 也对应 VAR xt，分别用 m1 存储原变量 xt 的均值，用 s1 存储原变量 xt 的标准偏差。

同理，MEAN＝m2 对应 VAR dy，STD＝s2 也对应 VAR dy，分别存储原变量 dy 的均值及标准偏差。

关于均值和标准偏差的分析与图 6.15 同。

已知用某种抗癌药物治疗 20 名中期患者一个月，测得疗效呈 2 的指数关系如下，求平均疗效。

疗效	2	4	8	16	32	64	128
人数	1	4	6	5	8	3	3

解：

- 由于疗效呈 2 的指数 2^x 关系，故设 $Y=\log_2(x)$。
- 设变量 L 为实验的人数。
- 计算出均值后，必须还原为 2^x。

程序内容见程序 6.14。

程序 6.14：

```
DATA ab1;
INPUT L x@@ ;
Y= log2(x);
CARDS;
1 2 4 4 6 8 5 16 8 32 3 64 3 128
;
PROC MEANS;
    VAR Y;
  FREQ L;
  OUTPUT OUT= KA MEAN= m;
DATA ab2;
    SET ab1;
    LX= 2* * Y; /* 还原为 2^Y * /
PROC PRINT DATA= ab2;
```

运行程序 6.14 产生图 6.17 所示的结果。

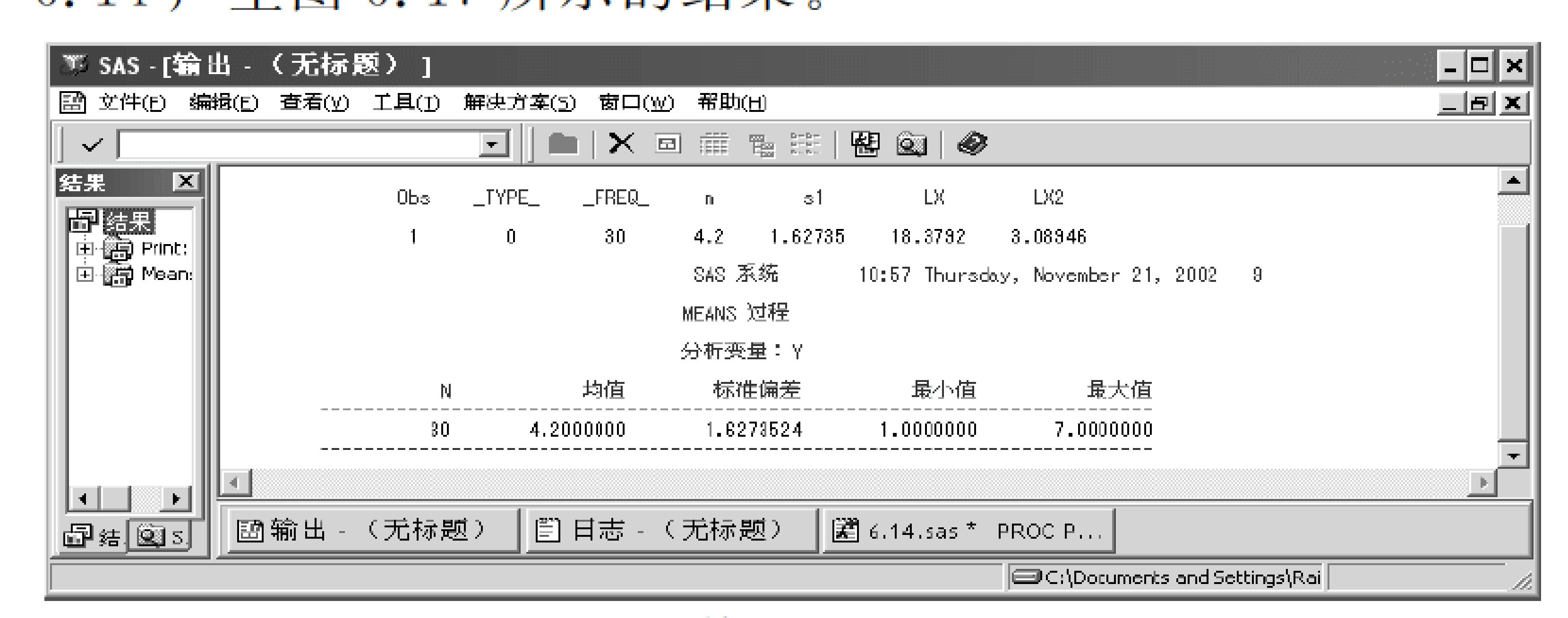


图 6.17　计算出均值后还原为 2^x

从图 6.17 看，平均疗效为 $2^{x}=2^{4.2}=18.3792$，标准偏差为 $2^{1.627352}=3.08946$。

6.7 用 PROC PLOT 过程画散点图

PROC PLOT 过程可产生双变量的散点图，从中可以看到双变量的相关程度（是否线性相关）。

例 13：画出低压与血糖的散点图，见程序 6.15。

程序 6.15：

```
DATA xt7;
LABEL location='地区' sex='性别' cy='抽烟量:支/日'
  mp='脉搏' dy='低压' xt='血糖';
INPUT sex location cy dy mp xt;
CARDS;
1 1 15 85 66 8.5
2 1 20 88 68 7.3
1 2 30 90 70 8.6
2 2 25 90 70 7.8
1 1 35 95 75 9.2
2 2 35 91 76 9.0
;
PROC SORT;BY sex;
  PROC PLOT;
  PLOT dy* xt/VAXIS=85 TO 95 BY 10;
```

运行程序 6.15 产生图 6.18 所示的结果。

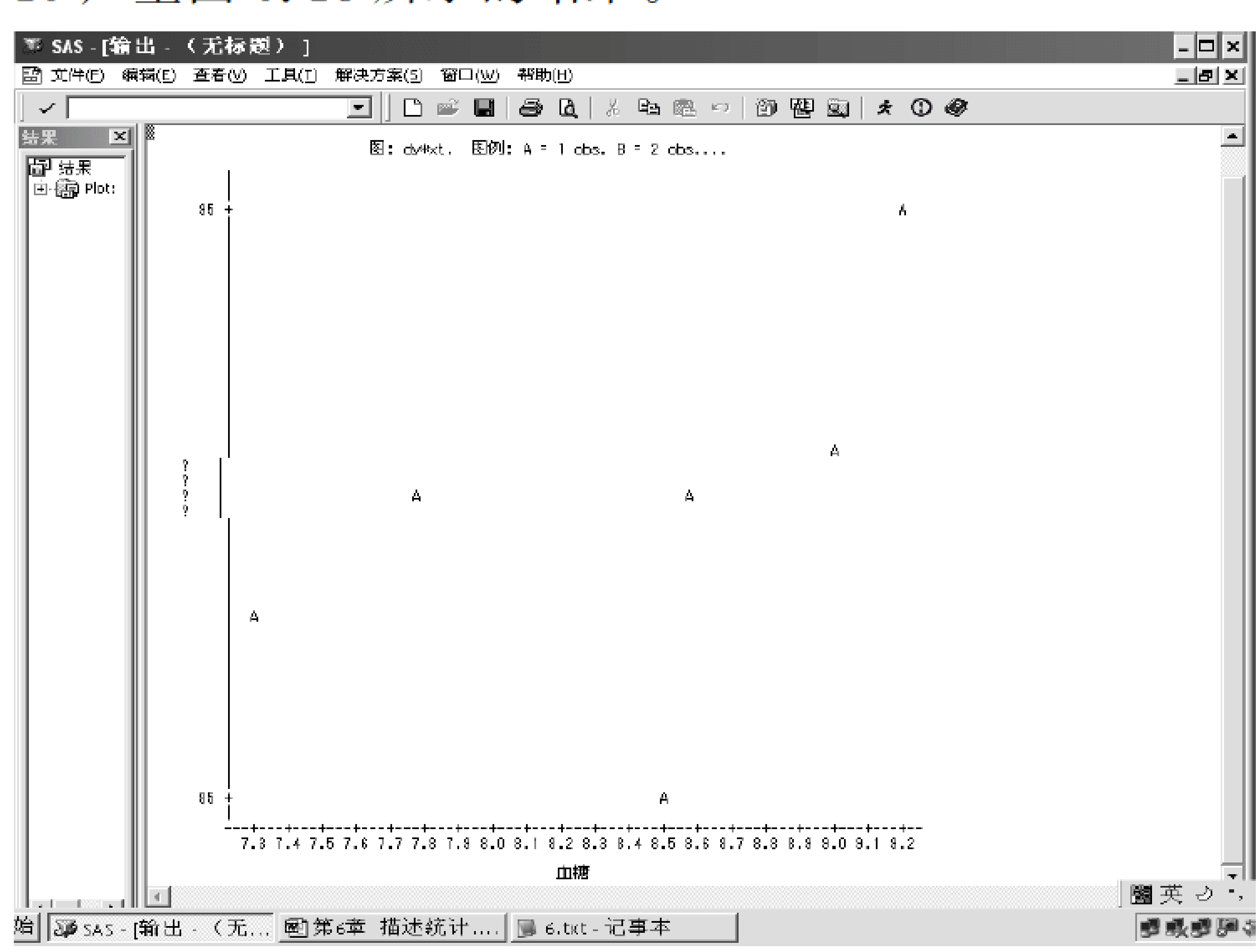


图 6.18 低压与血糖的散点图

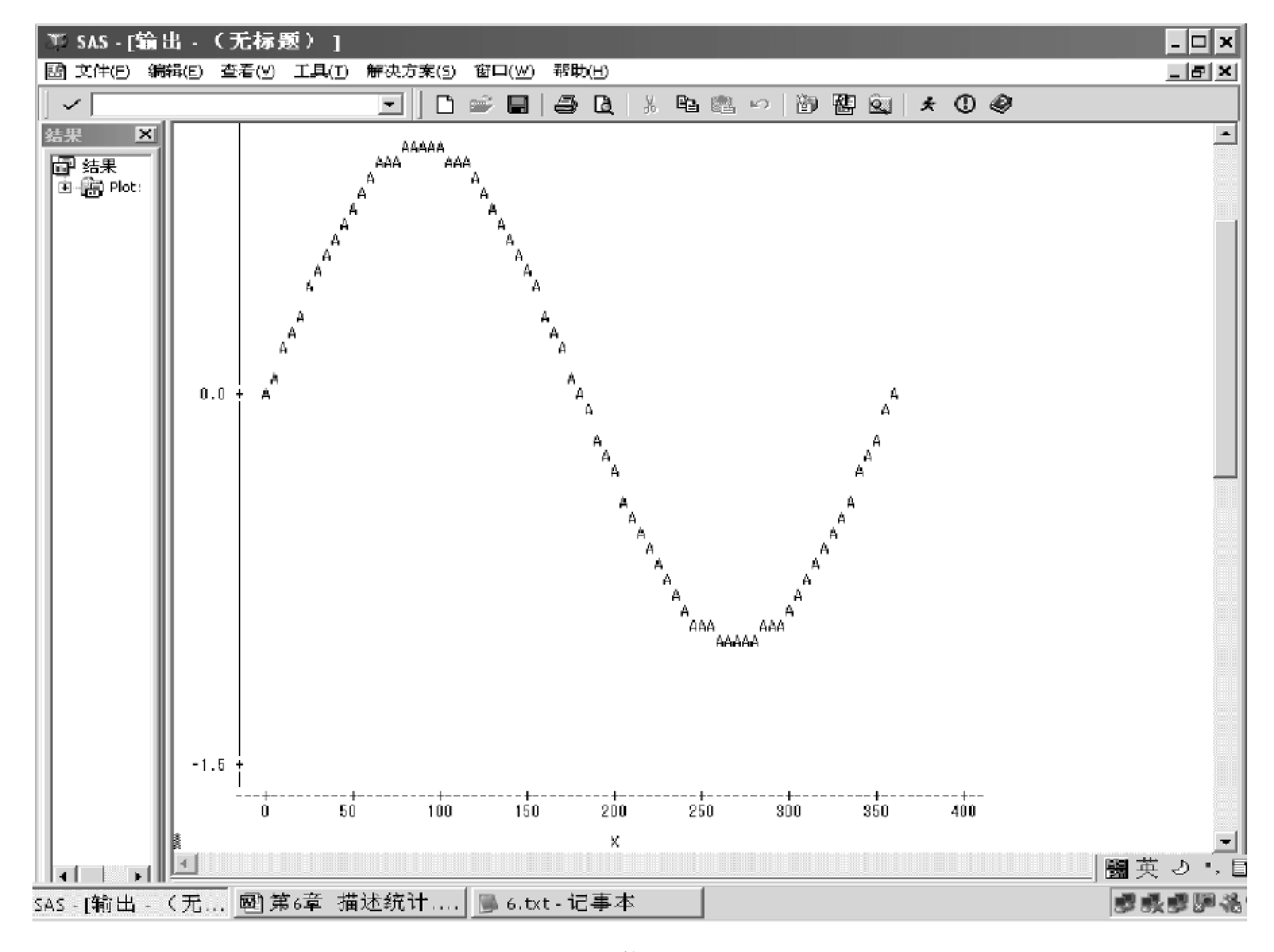
从图 6.18 看,低压与血糖的散点图没有呈现线性关系。低压 85 与血糖 8.5 的交叉点偏离其他点甚远。

例 14:PLOT 命令还可画出正弦和余弦图,见程序 6.16。

程序 6.16:

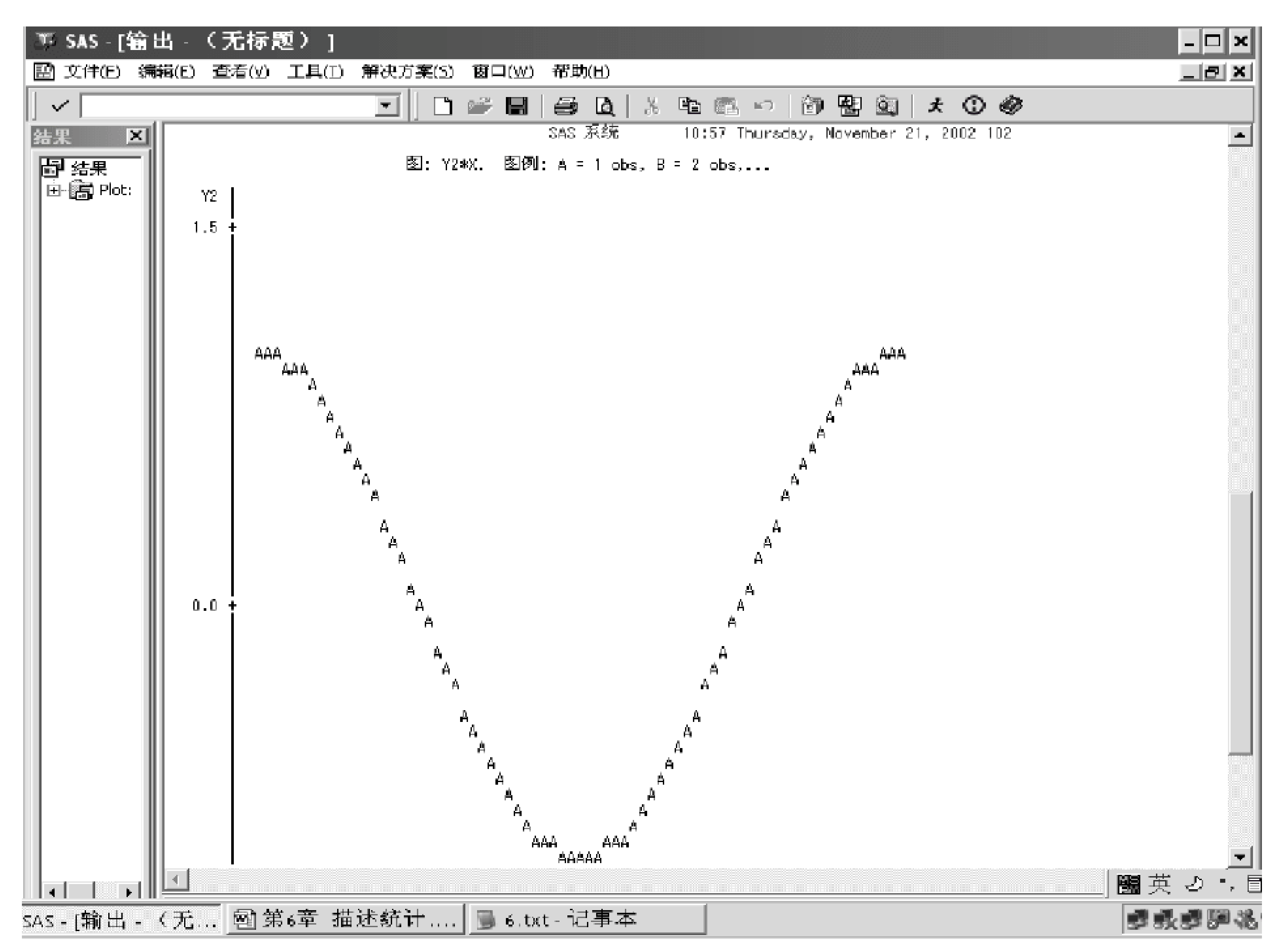
```
DATA;
DO X=0 TO 360 BY 5;
    Y1=SIN(x * 3.14159/180);
    Y2=COS(x * 3.14159/180);
    OUTPUT;
END;
  PROC PLOT;
 PLOT y1 * x;
 PLOT y2 * x;
```

运行程序 6.16 产生图 6.19 所示的结果。



(a) 正弦图

图 6.19 正弦和余弦图



(b) 余弦图

图 6.19（续）

6.8 用 RANK 过程进行非参数检验

对于标称(定类)数据、次序(定序)数据,以及从非正态分布的总体数据中抽取的区间(定距)和比例(定比)数据,由于不能采用参数检验,所以只能通过是否处于形状分布、均匀分布的观察来进行非参数检验。这时是对原始数据做秩分变换,即对原始数据从大到小(或从小到大)排列,进而赋予序号成为秩分。

6.8.1 什么是秩分

如表 6.1 所示,是人体中葡萄糖含量的秩分(变量 Rglucose)数据。原始数据 105 的秩分为 33.5,原始数据 95 的秩分为 14.0 等。

表 6.1 葡萄糖含量的秩分

OBS	Glucose	Rglucose
1	105	33.5
2	95	14.0
3	93	10.5
4	91	7.5
5	96	17.0

6.8.2 RANK过程命令

1. 命令格式

```
PROC RANK DATA=已建的数据集;
  TIES=MEAN|HIGH|LOW;     /* 3种取秩法,默认为 TIES=MEAN */
  DESCENDIN               /* 从大到小排列取秩,默认为升序取秩 */
  GROUP=n                 /* n为组数。计算分位数秩分。秩分为 0~(n-1)。
                             每组观察值可以相等或不等。当 n=100时产生百分位数,当 n=4时产生四
                             分位数 */
  FRACTION                /* 要求计算分数(小数)秩分,小数秩分=秩分/有效项数 n。
                             当指定 TIES=HIGH或省略 TIES项时,小数秩为右连续经验累计分布函数 */
  P|PERCENT               /* 计算百分数秩=(秩分/有效项数 n)*100% ,前提是有
                             TIES=HIGH选项 */
  NORMAL BLOW TUKEY VW    /* 先计算秩,后求正态分布。这些正态秩分接近于正态
                             分布所期望的顺序统计量 */
  SAVAGE                  /* 由秩分计算出 SAVAGE得分(指数得分) */
  OUT=数据集;             /* 默认为 OUT=_DATA_。输出数据集里含有输入数据集里的全部变量,加上由
                             RANK语句中的变量。如无 VAR语句,则输出数据集里含有全部数字型的变
                             量 */
```

例 15:

```
PROC RANK DATA=xt TIES=HIGH GROUPS=10 NORMAL=VW OUT=OU1;
    VAR dy xt mp;     /* 指定编秩的变量 */
    RANK R1 R2 R3;    /* 与"VAR dy xt mp;"中的 3个变量呼应,命名秩分名称 */
    [BY SEX;]         /* 按指定的变量(如 sex)分组取秩,但数据应该先用"BY sex;"语句取秩 */
```

2. 3种取秩法参阅表6.2

表6.2 3种取秩法

CALCIUM	TIES=MEAN	TIES=HIGH	TIES=LOW
10.1	8.0	10	6
8.7	6.5	6	7
10.5	9.0	10	8
10.6	10.0	10	10

6.8.3 秩分计算

1. 正态得分

RANK具有3种正态得分计算法:

BLOW :$Z_i=\Psi(R_i-3/8)/(n+1/4)$

TUKEY：Zi＝Ψ(Ri－1/3)/(n＋1/3)

VW ：Zi＝Ψ(Ri)/(n＋1)

其中：Ψ为逆累积正态概率函数。Ri为第 i 个秩。n 为“VAR 语句”中变量的有效个案数目。VW 为 Van Der Wacerden 的缩写，其得分用于非参数定位检验。

2. 分位秩的算法为 FLOOR(RANK＊G/(n＋1))

式中，RANK 为数据的秩分，G 为 GROUP＝v 的组数，n 为“VAR 语句”中变量的有效个案数目。

3. 指数得分(SAVAGE 得分)

对秩次(秩分)进行指数变换的公式如下：

$$Yi = \sum_{J=n-Ri+1}^{n} (1/j) - 1$$

式中，Ri 为第 i 个秩。当原始数据为指数分布时，先进行指数得分变换，再进行统计分析。

6.8.4 运用举例

当数据变换为秩分后，一般可与 PLOT、NPAR1WAY、MEANS 等过程进行联合分析。

1. 用秩分检验数据的正态性

利用 RANK 过程先计算正态得分和指数得分，再检验这些得分是否服从正态分布或指数分布，即把正态得分作为横轴，原始数据作为纵轴，用 PLOT 过程画图，如果图形的图点呈现一直线，则为正态分布。

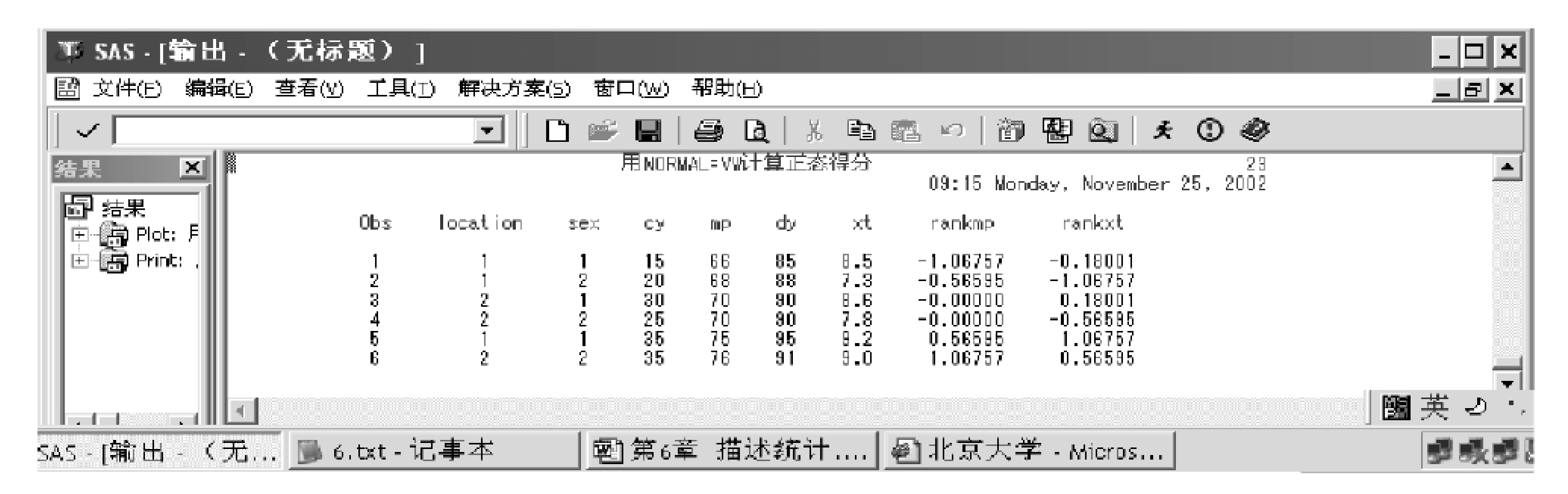
例 16：检验血糖数据是否服从正态分布，见程序 6.17。

程序 6.17：

```
DATA xt;
LABEL location='地区' sex='性别' cy='抽烟量:支' mp='脉搏'
                  dy='低压' xt='血糖';
INPUT id sex location cy dy mp xt;
CARDS;
1 1 15 85 66 8.5
2 1 20 88 68 7.3
1 2 30 90 70 8.6
2 2 25 90 70 7.8
1 1 35 95 75 9.2
2 2 35 91 76 9.0
;
PROC RANK NORMAL=VW OUT=OU2;
```

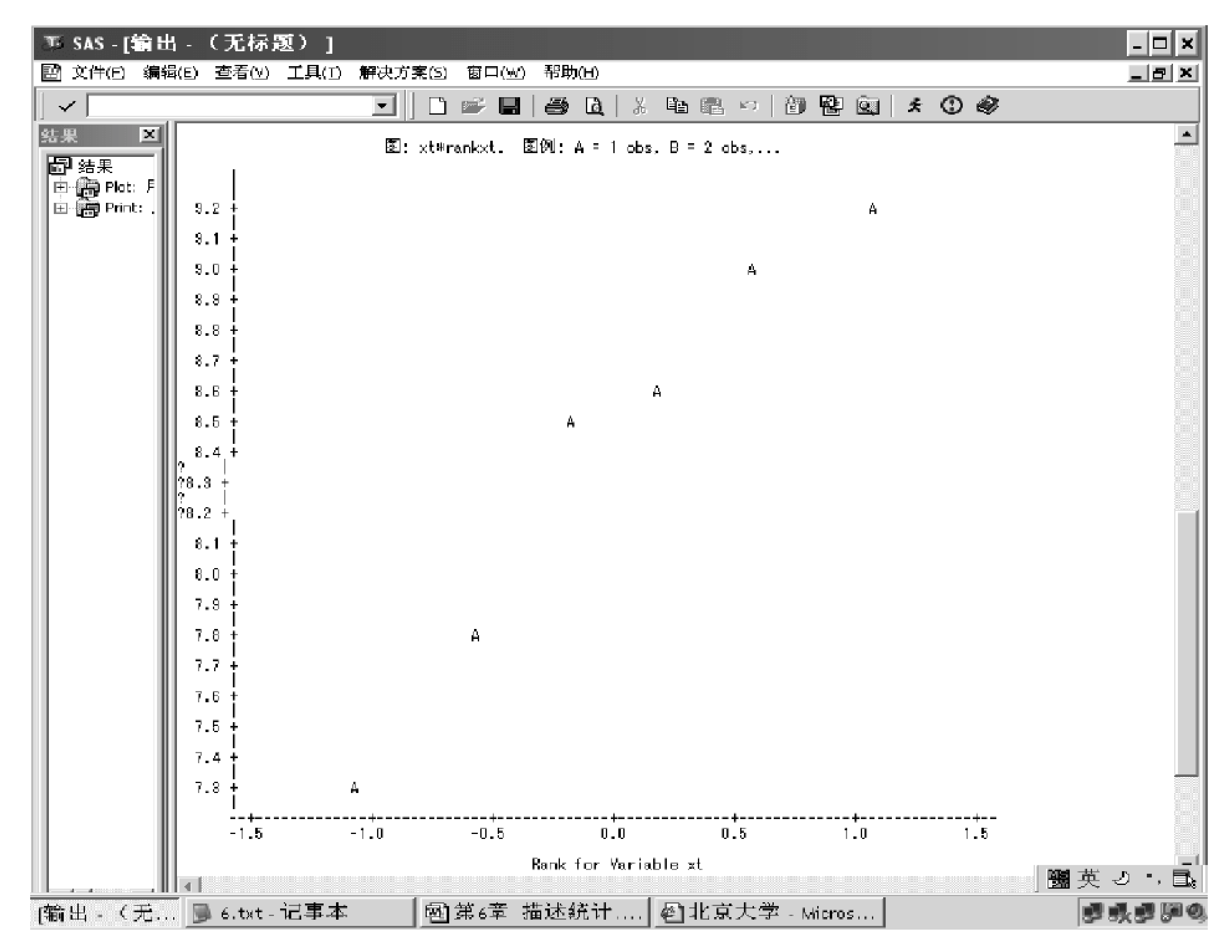
```
        VAR mp xt;
        RANKS rankmp rankxt;
PROC PRINT DATA=ou2;
TITLE '用 NORMAL=VW计算正态得分';
PROC PLOT;
        PLOT mp * rankmp xt * rankxt;
```

运行程序 6.17 产生图 6.20 所示的结果。



Obs	location	sex	cy	mp	dy	xt	rankmp	rankxt
1	1	1	15	66	85	8.5	-1.06757	-0.18001
2	1	2	20	68	88	7.3	-0.56595	-1.06757
3	2	1	30	70	90	8.6	-0.00000	0.18001
4	2	2	25	70	90	7.8	-0.00000	-0.56595
5	1	1	35	75	95	9.2	0.56595	1.06757
6	2	2	35	76	91	9.0	1.06757	0.56595

(a) 计算正态得分



(b) xt*rankxt的散点图

图 6.20　计算正态得分及散点图

从图 6.20(a)看，mp 数据的秩分为 rankmp，xt 数据的秩分为 rankxt，分别已生成。再从图 6.20(b)看，原始数据 xt 与秩分 rankxt 的散点图基本呈现一直线，说明血糖数据

基本上趋于正态分布。

2. 秩和检验

(1) 双样本秩和检验

RANK先将原始数据变为秩分，然后用NPAR1WAY过程中的WILCOXON选项对秩分进行t检验，进而获得双样本秩和检验。

例17：单因素双样本秩和检验，见程序6.18。

程序6.18：承接程序6.3的数据。

```
DATA xt;
LABEL location='地区' sex='性别' cy='抽烟量:支' mp='脉搏'
                  dy='低压' xt='血糖';
INPUT sex location cy dy mp xt;
CARDS;
1 1 15 85 66 8.5
2 1 20 88 68 7.3
1 2 30 90 70 8.6
2 2 25 90 70 7.8
1 1 35 95 75 9.2
2 2 35 91 76 9.0
;
PROC RANK NORMAL=VW OUT=OU2;
    VAR mp xt;
    RANKS rankmp rankxt;
PROC NPAR1WAY WILCOXON;
    CLASS sex;
     VAR rankxt; /* 单因素双样本秩和检验 * /
```

运行程序6.18产生图6.21所示的结果。

从图6.21看，双侧概率(Two-Sided Pr>|Z|)为0.3827，单侧概率(One-Sided Pr>Z)为0.1914，都不显著。说明秩和趋于正态分布。

双样本的秩和(Sum of Scores)分别为13.0及8.0，不相等。

(2) 多样本秩和检验

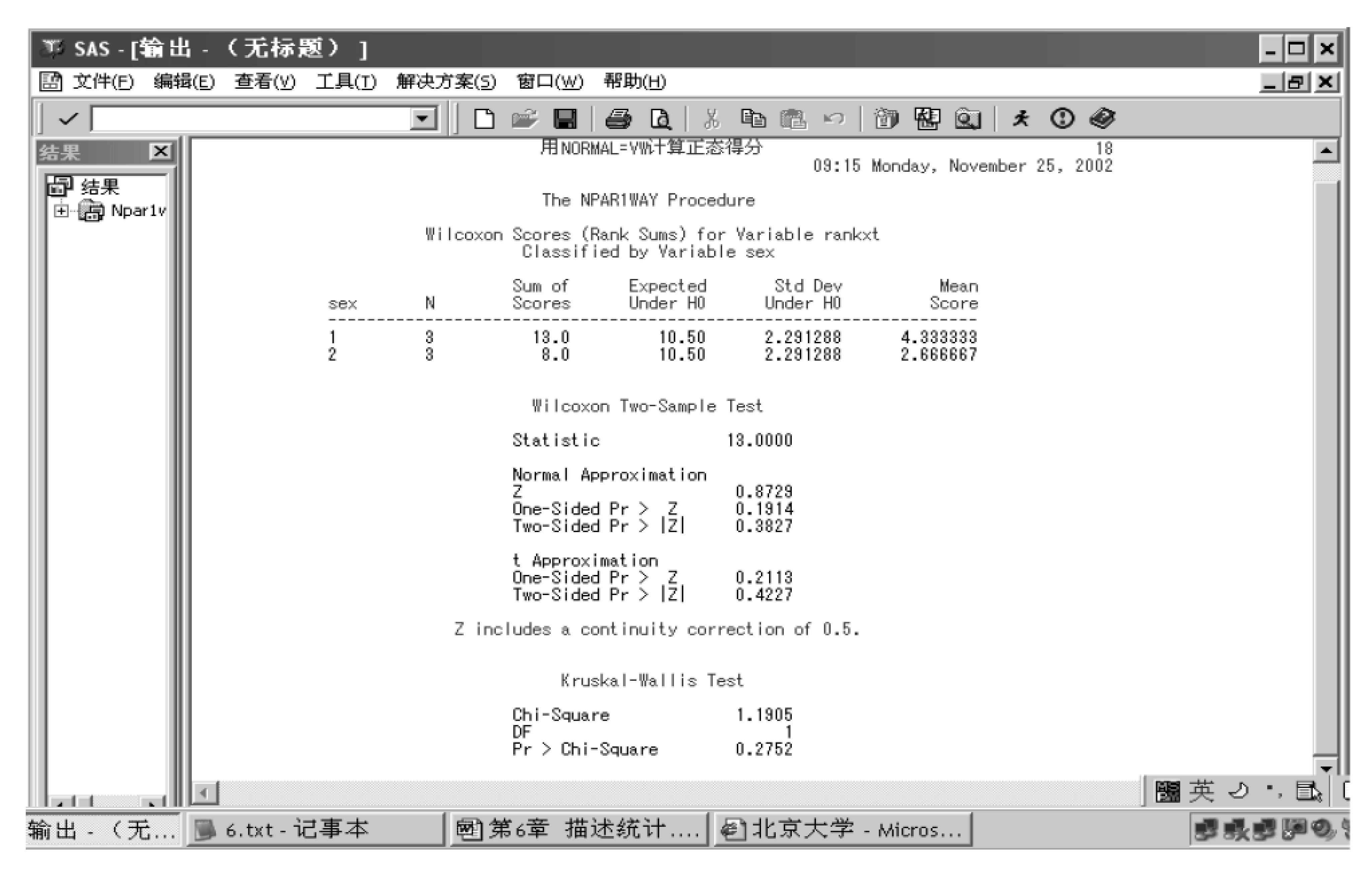
多样本秩和检验是将原始数据的秩分(秩次)进行多因素(多自变量)多水平的方差分析，即Kruskal-Walis检验。

检验法：仍用NPAR1WAY过程中的WILCOXON选项对秩分进行t检验，进而获得多样本秩和检验。

输出结果：只输出卡方及P值，比双样本秩和检验少一个Z值。

3. 等级相关分析

斯皮尔曼(Spearman)等级相关分析实质上是秩相关分析。当变量分布是次序的，或非正态的，或分布为未知时，可采用这种检验法。



```
用NORMAL=VW计算正态得分                                   18
                                   09:15 Monday, November 25, 2002

                The NPAR1WAY Procedure

     Wilcoxon Scores (Rank Sums) for Variable rankxt
              Classified by Variable sex

                   Sum of     Expected     Std Dev        Mean
sex        N       Scores     Under H0     Under H0       Score
------------------------------------------------------------------
1          3         13.0        10.50     2.291288    4.333333
2          3          8.0        10.50     2.291288    2.666667

              Wilcoxon Two-Sample Test

              Statistic              13.0000

              Normal Approximation
              Z                       0.8729
              One-Sided Pr >  Z       0.1914
              Two-Sided Pr > |Z|      0.3827

              t Approximation
              One-Sided Pr >  Z       0.2113
              Two-Sided Pr > |Z|      0.4227

        Z includes a continuity correction of 0.5.

                 Kruskal-Wallis Test

              Chi-Square              1.1905
              DF                           1
              Pr > Chi-Square         0.2752
```

图 6.21 单因素双样本秩和检验

等级相关分析的过程命令为：

```
PROC CORR SPEARMAN;
```

程序 6.19：

```
DATA c1;
LABEL location='地区' sex='性别' cy='抽烟量:支' mp='脉搏'
                  dy='低压' xt='血糖';
INPUT sex location cy dy mp xt;
CARDS;
1 1 15 85 66 8.5
2 1 20 88 68 7.3
1 2 30 90 70 8.6
2 2 25 90 70 7.8
1 1 35 95 75 9.2
2 2 35 91 76 9.0
;
PROC CORR SPEARMAN; /* 斯皮尔曼(Spearman)等级相关分析 */
   VAR cy xt;
PARTIAL sex;
```

运行程序 6.19 产生图 6.22 所示的结果。

从图 6.22 看，抽烟与血糖的斯皮尔曼等级相关系数为 0.9299，其显著性水平为 0.022(小于 α 值 0.05)很显著。说明抽烟多了会致使血糖很快增高。

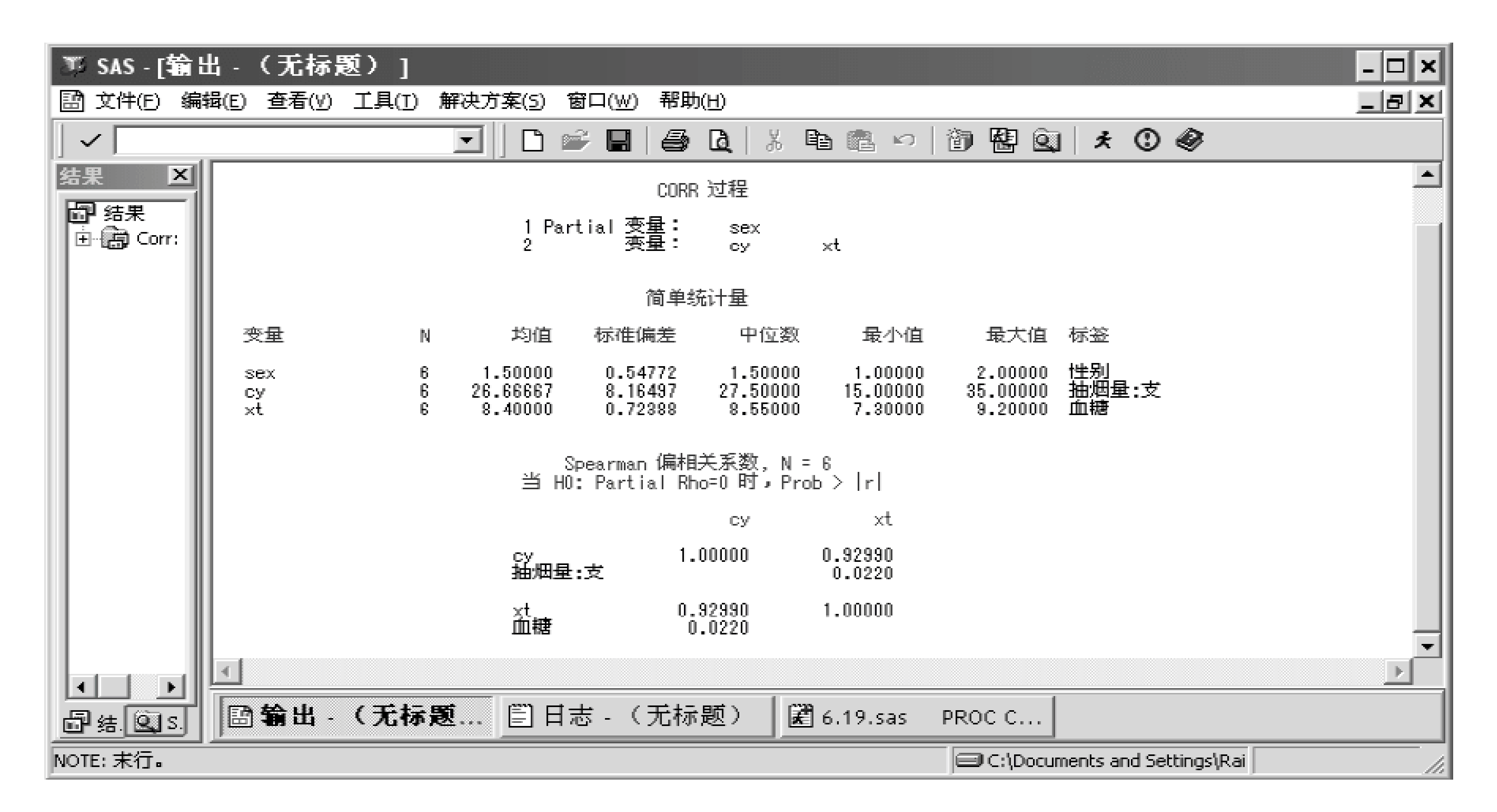


图 6.22 斯皮尔曼(Spearman)等级相关分析

4. 利用秩分计算秩和比

对于一些世界性的健康指标(如出生率、死亡率、婴儿死亡率,0 岁、1 岁、65 岁的期望寿命以及结婚率等),可先排序然后对秩求和。最后将和数除以(指标个数 * 个案数),即:

$$RSR = \left(\sum R\right) \Big/ (m * n)$$

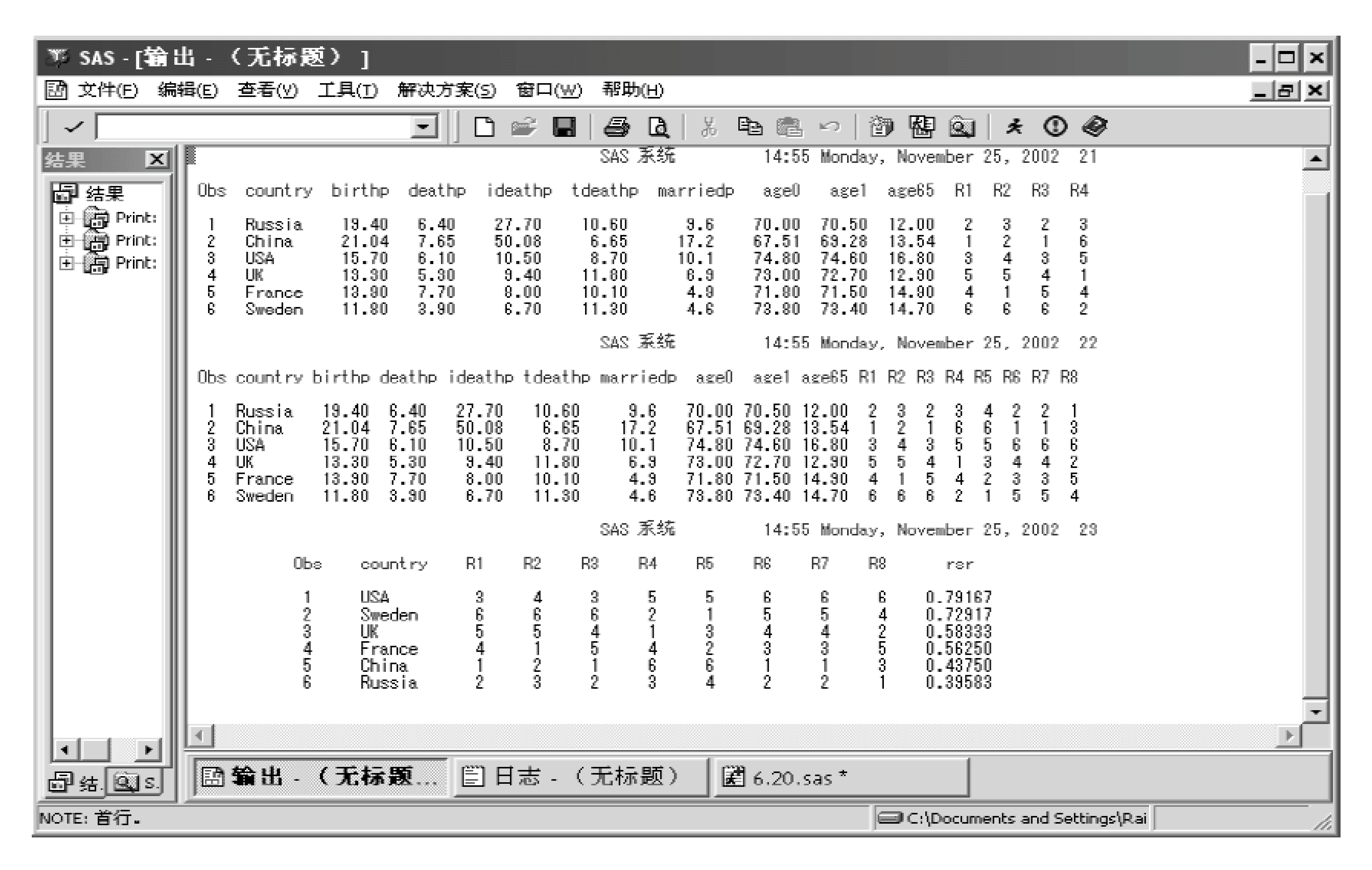
式中,RSR 既是秩和比,又是健康指数,例如:计算健康指数,见程序 6.20。

程序 6.20:

```
DATA rsr1;
INPUT country $  birthp deathp ideathp tdeathp marriedp age0 age1 age65;
LABEL country= '国家' birthp= '出生率' deathp= '死亡率'
      ideathp= '婴儿死亡率'
                 tdeathp= '总死亡率' marriedp= '结婚率' age0= '0 岁期望寿命'
                 age1= '1 岁期望寿命' age65= '65 岁期望寿命';
CARDS;
Russia   19.40   6.40   27.70   10.60    9.60   70.00   70.50   12.00
China    21.04   7.65   50.08    6.65   17.20   67.51   69.28   13.54
USA      15.70   6.10   10.50    8.70   10.10   74.80   74.60   16.80
UK       13.30   5.30    9.40   11.80    6.90   73.00   72.70   12.90
France   13.90   7.70    8.00   10.10    4.90   71.80   71.50   14.90
Sweden   11.80   3.90    6.70   11.30    4.60   73.80   73.40   14.70
;
PROC RANK DESCENDING OUT= OUTP1; /* 出生率、死亡率等指标应降排序,越低越好 */
    VAR birthp deathp ideathp tdeathp;
    RANKS R1-R4;
PROC RANK OUT= OUTP2;           /* 结婚率和期望率等指标应升排序,越高越好 */
```

```
    VAR marriedp age0 age1 age65;
    RANKS R5-R8;
DATA rsr2;                    /* 建立第 2 个输入数据集"秩和比" */
    MERGE OUTP1 OUTP2;        /* 将数据集 rsr1 和 rsr2 连接在一起 */
    rsr=SUM(OF R1-R8)/(6*8);  /* 计算秩和比 */
    KEEP country R1-R8 rsr;   /* 输出数据集里所保留的变量 */
PROC SORT;
    BY DESCENDING rsr;
PROC PRINT;
RUN;
```

运行程序 6.20 产生图 6.23 所示的结果。



SAS 系统　　14:55 Monday, November 25, 2002　21

Obs	country	birthp	deathp	ideathp	tdeathp	marriedp	age0	age1	age65	R1	R2	R3	R4
1	Russia	19.40	6.40	27.70	10.60	9.6	70.00	70.50	12.00	2	3	2	3
2	China	21.04	7.65	50.08	6.65	17.2	67.51	69.28	13.54	1	2	1	6
3	USA	15.70	6.10	10.50	8.70	10.1	74.80	74.60	16.80	3	4	3	5
4	UK	13.30	5.30	9.40	11.80	6.9	73.00	72.70	12.90	5	5	4	1
5	France	13.90	7.70	8.00	10.10	4.9	71.80	71.50	14.90	4	1	5	4
6	Sweden	11.80	3.90	6.70	11.30	4.6	73.80	73.40	14.70	6	6	6	2

SAS 系统　　14:55 Monday, November 25, 2002　22

Obs	country	birthp	deathp	ideathp	tdeathp	marriedp	age0	age1	age65	R1	R2	R3	R4	R5	R6	R7	R8
1	Russia	19.40	6.40	27.70	10.60	9.6	70.00	70.50	12.00	2	3	2	3	4	2	2	1
2	China	21.04	7.65	50.08	6.65	17.2	67.51	69.28	13.54	1	2	1	6	6	1	1	3
3	USA	15.70	6.10	10.50	8.70	10.1	74.80	74.60	16.80	3	4	3	5	5	6	6	6
4	UK	13.30	5.30	9.40	11.80	6.9	73.00	72.70	12.90	5	5	4	1	3	4	4	2
5	France	13.90	7.70	8.00	10.10	4.9	71.80	71.50	14.90	4	1	5	4	2	3	3	5
6	Sweden	11.80	3.90	6.70	11.30	4.6	73.80	73.40	14.70	6	6	6	2	1	5	5	4

SAS 系统　　14:55 Monday, November 25, 2002　23

Obs	country	R1	R2	R3	R4	R5	R6	R7	R8	rsr
1	USA	3	4	3	5	5	6	6	6	0.79167
2	Sweden	6	6	6	2	1	5	5	4	0.72917
3	UK	5	5	4	1	3	4	4	2	0.58333
4	France	4	1	5	4	2	3	3	5	0.56250
5	China	1	2	1	6	6	1	1	3	0.43750
6	Russia	2	3	2	3	4	2	2	1	0.39583

图 6.23　健康指数

结果分析：

从图 6.23 看，我国的健康指数曾经为第五名，美国第一。

习　题　6

1. PROC FREQ 过程可以做哪两种频率表？
2. 试计算 sex 变量的一维频数分布，并把缺失值当作有效值统计。
3. 试计算"定类-定类"双变量交叉汇总与结合测量。
4. 试计算"定比-定比"双变量交叉汇总与结合测量。
5. 试用 PROC CHART 过程中的水平直方图描述血糖数据。
6. 什么情况下要采用 RANK 过程进行非参数检验？
7. 试对低压变量 dy 进行单因素双样本秩和检验。

第 7 章 chapter 7

均值比较与 T 检验

本章介绍常用的两个样本(两个子总体)均值的比较和 T 检验。

7.1 均值比较的方法

均值比较(如表 7.1 所示)是教学科研中常用的一种统计分析法,它有以下几种比较法。

表 7.1 两个均值的比较

样　本	比　较　法
样本均值与总体均值的比较	DATA xt7;INPUT x @@; Y=x-假定值; PROC MEANS MEAN STD T PRT; VAR Y;
AB 两个样本的均值比较	PROC MEANS MEAN STD STDERR T PRT; VAR A B;
两组(男女,或对照组与实验组)的均值比较	PROC TTEST; CLASS GROUP; VAR x;
非参数检验: 两组(男女,或对照组与实验组)样本呈现非正态分布,或分布不确定,或为偏态分布时,采用非参数检验	PROC NPAR1WAY WILCOXON; CLASS GROUP; VAR x;

7.1.1 配对样本的均值比较

配对样本的均值比较分为以下两种:

(1) 样本均值与总体均值之间的比较,以及均值差异的显著性检验。其方法是:先用 MEANS 过程计算两个均值之差,再用 MEANS 过程中的 t 统计量进行均值差异的显著性检验。

(2) 同一对象在实验前后的结果(即均值)比较,或配对样本的均值比较。检验过程

同上。

7.1.2 两个独立样本的均值差检验

两个独立样本的均值差检验分为以下两种：

(1) 男女两组(或对照组与实验组)的均值比较。

(2) 甲乙两组平均健康指标(如血压、血糖、身高、体重)的均值比较。

两个独立样本(两组)的均值差检验，要求两组数据各自独立、且来自服从正态分布的总体。

检验法：采用 TTEST 过程。

① 先检验两个独立样本(两组)的方差是否相等。其原假设为：

H0：两个独立样本(两组)的方差相等。

② 再检验两个独立样本(两组)的均值是否相等。

H0：两个独立样本(两组)的均值相等。

7.2 MEANS 过程及其 t 统计量

MEANS 过程是计算两个样本均值的差异，MEANS 过程中的 t 统计量是对均值差异的显著性检验。

1. MEANS 过程的命令格式

```
PROC MEANS MEAN STD STDERR T PRT;
    VAR v1 v2;
```

2. 格式说明

PROC MEANS 后面有 5 个默认的统计量：MEAN(均值)、STD(均值的标准差)、STDERR(标准误差)、T(均值差异的显著性检验的统计量 t)、PRT(统计量 t 的概率，此值要与 α 值 0.05 进行比较)。

VAR v1 v2：进行均值比较的数字型变量 v1 v2 等。

3. 举例

例 1：样本均值与总体均值之间的比较，以及均值差异的显著性检验。设每人体内血糖的标准含量为 5。现对 5 名受视者空腹抽取 5 次血样测得血糖含量如下：

```
5.0  4.9  5.1  4.8  5.3
```

试检验其平均血糖与标准均值 5 有无显著差异。命令语句见程序 7.1。

程序 7.1：

```
DATA xt7;
```

```
INPUT v @@;
    Y=v-5; /* 血糖含量与标准值 5的差值 */
CARDS;
5.0  4.9  5.1  4.8  5.3
PROC MEANS MEAN STD T PRT; /* 计算最主要的统计量 */
    VAR Y;
```

运行程序 7.1 产生图 7.1 所示的结果。



图 7.1 平均血糖与标准值的差异检验

H0：两个均值之差为 0。

从图 7.1 看，均值差为 0.02，标准偏差为 0.1923538。t 值 0.23 小，t 的概率值 0.8276 大于 α 值 0.05。

所以没有足够的理由拒绝 H0，说明平均血糖与标准值 5 很接近。

例 2：配对样本 t 检验。对 20 位肿瘤患者，其中的 10 个人采用药物＋化疗治疗，另 10 个人采用药物＋放疗治疗。两周后测得体重增加(单位：公斤)见表 7.2。试做两种疗效差异性检验。

表 7.2 20 位肿瘤患者体重增加(单位：公斤)数据

配对个案	1	2	3	4	5	6	7	8	9	10
药物＋化疗	0.50	0.75	0.80	0.91	0.69	0.48	0.33	0.66	0.51	0.59
药物＋放疗	0.60	0.65	0.78	0.82	0.56	0.23	0.40	0.51	0.47	0.48

解法：见程序 7.2。

程序 7.2：

```
DATA TZ;
INPUT v1 v2;
  D=v1-v2;                    /* 20位肿瘤患者体重之差 */
LABEL v1='药物+化疗' v2='药物+放疗';
CARDS;
0.50 0.75 0.80 0.91 0.69 0.48 0.33 0.66 0.51 0.59
0.60 0.65 0.78 0.82 0.56 0.23 0.40 0.51 0.47 0.48
PROC MEANS MEAN STD T PRT;    /* 计算最主要的统计量 */
    VAR D;
RUN;
```

运行程序 7.2 产生图 7.2 所示的结果。

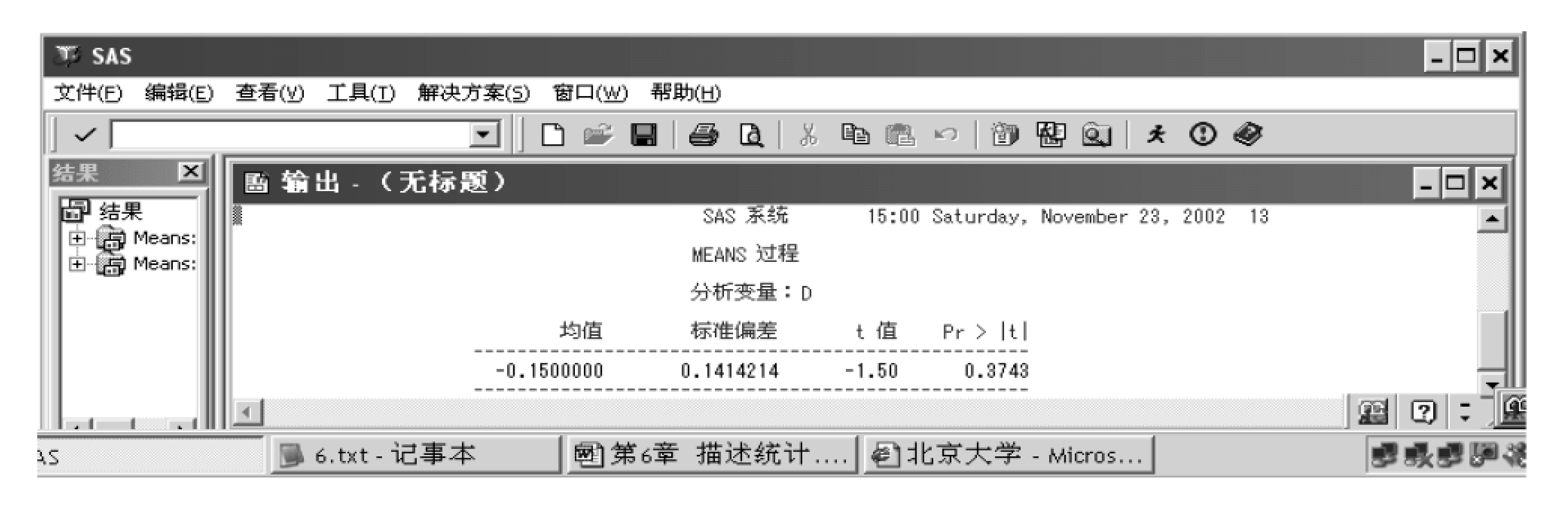


图 7.2 20 位肿瘤患者体重之差

结果分析：

H0：两个均值之差为 0。

从图 7.2 看，均值差为 0.15，标准偏差为 0.1414。t 值 1.50 小，t 的概率值 0.3743 大于 α 值 0.05。

所以没有足够的理由拒绝 H0，说明两种疗效很接近。

例 3：自身在治疗前后血小板均值差的检验。

对 10 位血小板偏低者进行药物治疗一个月后测得血小板数据见表 7.3。问该药是否能提高血小板的水平。

表 7.3 10 位血小板偏低者药物治疗一个月后血小板数据

治疗前后	1	2	3	4	5	6	7	8	9	10
治疗前	8.5	9.1	8.0	7.6	6.9	7.7	8.0	8.3	9.1	8.2
治疗后	10.3	9.3	10.1	8.9	8.2	9.1	8.5	9.0	10.1	9.2

解法：见程序 7.3。

程序 7.3：

```
DATA xxb;
INPUT v1 v2 @@;
  D=v2-v1;                    /* 10位被试者血小板数据之差 */
CARDS;
 8.5  9.1  8.0  7.6  6.9  7.7  8.0  8.3  9.1  8.2
10.3  9.3  10.1  8.9  8.2  9.1  8.5  9.0  10.1  9.2
PROC MEANS MEAN STD T PRT;    /* 计算最主要的统计量 */
    VAR D;
RUN;
```

运行程序 7.3 产生图 7.3 所示的结果。

结果分析：

H0：两个均值之差为 0。

从图 7.3 看，均值差为 0.13，标准偏差为 0.8300602。但 t 值 0.50 小，t 的概率值

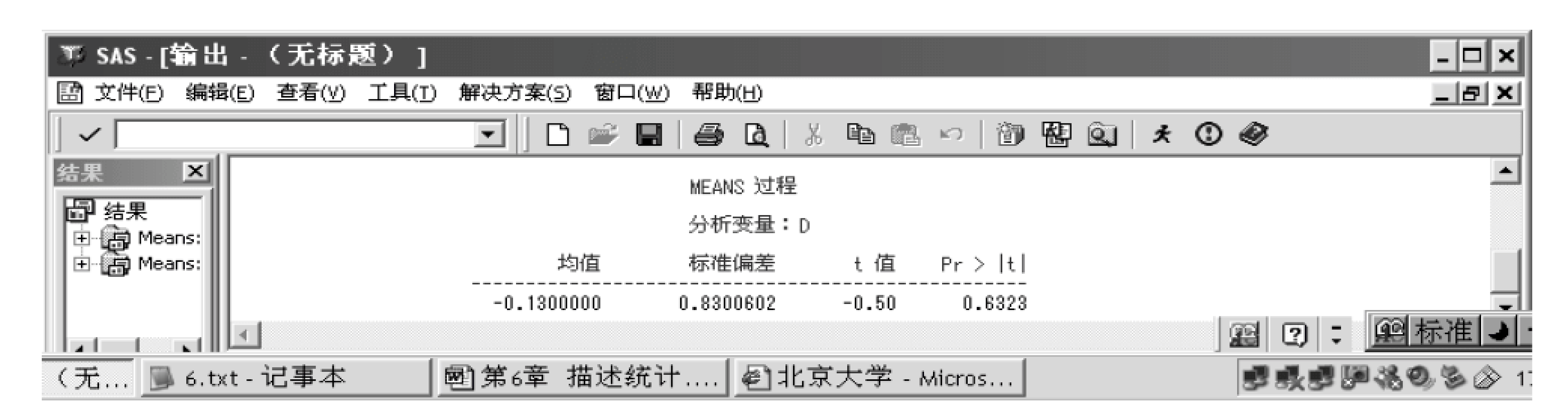


图 7.3　10 位被试者血小板数据之差

0.6323 大于 α 值 0.05。

所以没有足够的理由拒绝 H0，说明治疗前后体重差异不显著。

7.3　TTEST 过程及其 t 检验

TTEST 过程适合于“两个独立样本的均值差为 0”的 t 检验。所谓两个样本互为独立，是指两组数据各自独立、并且是来自服从正态分布的总体中的两个样本。

例如有男女两个样本，男性与女性之间互不相关，此外，男女人数可以不等。

例 4：将被试者分为对照组和实验组两组，使用同一种抗癌药物，一个月后测得肿块大小见表 7.4 所示。试做两组疗效差异性的检验。命令语句见程序 7.4a。

表 7.4　对照组和实验组肿块数据

实验组	1	48	70	60	50	65	68	75	88	64	58
对照组	2	40	80	75	48	70	60	70	60	70	62

解法：见程序 7.4a。

程序 7.4a：

```
DATA dbs;
INPUT group x @@;
CARDS;
1 48 1 70 1 60 1 50 1 65 1 68 1 75 1 88 1 64 1 58
2 40 2 80 2 75 2 48 2 70 2 60 2 70 2 60 2 70 2 62
;
PROC TTEST;
CLASS group;
VAR x;
RUN;
```

程序 7.4a 说明：

程序 7.4a 中用 CLASS 语句指定 group 为分组变量，表示两组被试者，其中 group=1 为实验组，group=2 为对照组。变量 x 表示肿瘤的变化量，必须是数字型变量。

运行程序 7.4a 产生图 7.4 所示的结果。

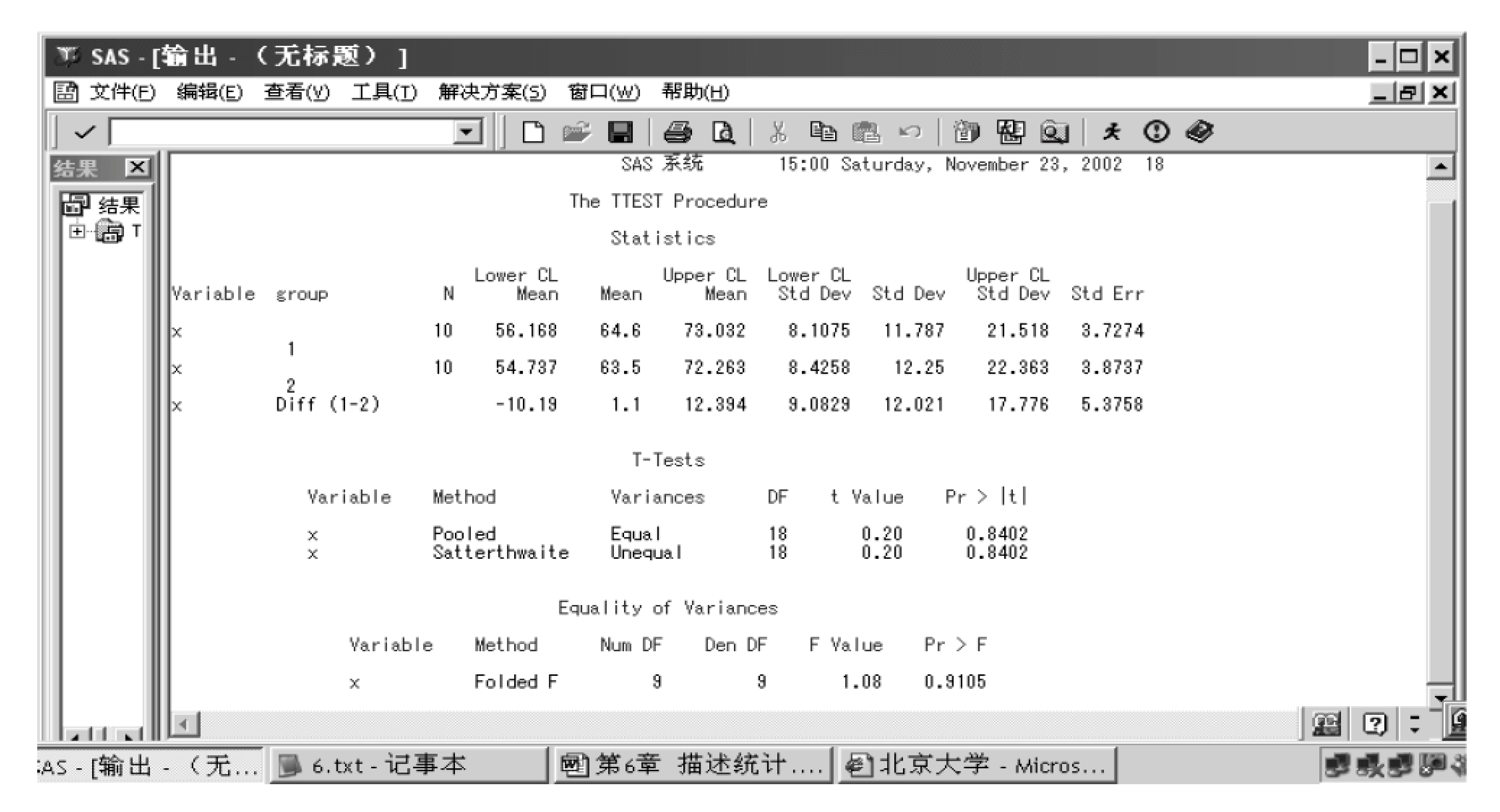


图 7.4　对照组和实验组肿块比较

结果分析：

(1) 方差相等的检验

H0：两个子总体(两个样本)的方差相等。

检验：

从图 7.4 看，F 值的概率“Pr>F”为 0.9105，此值大于 α 值 0.05，所以没有足够的理由拒绝 H0，说明两个子总体(两个样本)的方差相等。

当方差相等时应该再观察图 7.4 中的 Equal 一行的 t 值及其显著性水平。

(2) 均值相等的检验

H0：两个子总体(两个样本)的均值相等。

检验：

从图 7.4 中的 Equal 一行的 t 值及其显著性水平 0.8402 看，大于 α 值 0.05，所以没有足够的理由拒绝 H0，说明两个子总体(两个样本)的均值差异不显著(注：不宜说均值相等)。

程序 7.4a 是采用常用的编程法。下面改用 DO…END 语句编程，见程序 7.4b。

程序 7.4b：

```
DATA dbs2;
  DO group=1 TO 2;          /* 读取外围 2组 */
     INPUT n;
     DO I=1 TO n;           /* 读取内围各 10个观察值 */
       INPUT x @@;
       OUTPUT;
     END;
  END;
```

```
  DORP I;
CARDS;
10
48 70 60 50 65 68 75 88 64 58
10
40 80 75 48 70 60 70 60 70 62
PROC  TTEST;
  CLASS group;
  VAR x;
RUN;
```

运行程序 7.4b,也可产生图 7.4 所示的结果。

7.4 非参数检验

非参数检验也是检验两个子总体均值的差异。但是,非参数检验是针对两个子总体的分布很难确定"正态分布"的情形,而且还针对已知两个子总体的分布趋于偏态分布的数据。

非参数检验采用 SAS 中的 NPAR1WAY 过程。它基于经验分布函数,计算出几个统计量来检验变量的分布在跨组时是否有"相同的位置参数"。这些统计量如下:

- Wilcoxon 得分
- 中位数得分
- Savage 得分
- VW(Van Der Wacerden)得分

7.4.1 用 NPAR1WAY 过程做非参数检验

NPAR1WAY 过程专门用于单因素非参数检验。

1. 过程命令格式

```
PROC NPAR1WAY WILCOXON;
    CLASS v;      /* 指定一个分类变量 */
    VAR v2;       /* 指定数字型因变量 */
```

2. 格式说明

选项中只有一个 WILCOXON 统计量,是对数据的秩分或 WILCOXON 得分进行分析。

对于单因素二水平,可进行 WILCOXON 秩和检验(U 检验)。对于单因素三水平及更多水平,可进行 Kruskal-Wallis 检验(H 检验)。

7.4.2 举例

例 5：用 WILCOXON 秩和检验对实验组与对照组此双样本(即二水平)均差的显著性检验，见程序 7.5。

程序 7.5：

```
DATA dbs3;
INPUT group x @ @ ;
CARDS;
1 48 1 70 1 60 1 50 1 65 1 68 1 75 1 64 1 66 1 58
2 60 2 80 2 75 2 48 2 70 2 60 2 70 2 60 2 70 2 72
;
PROC NPAR1WAY WILCOXON;
     CLASS group;          /* 指定 group 为分类变量 * /
     VAR x;                /* 指定数字型因变量 x* /
RUN;
```

运行程序 7.5 产生图 7.5 所示的结果。

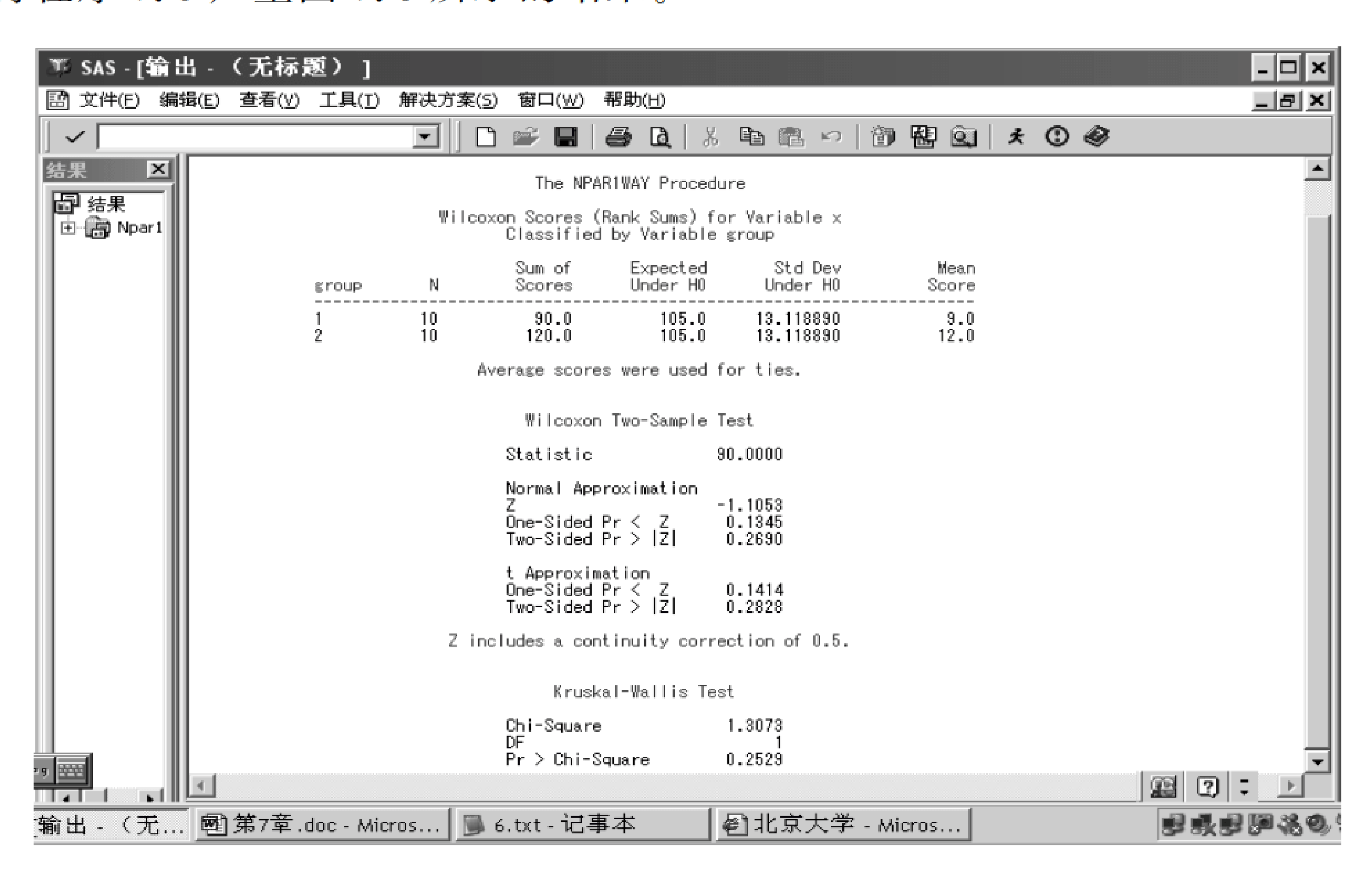


图 7.5 双样本均差的显著性检验

结果分析：

H0：实验组与对照组消肿效果没有差异。

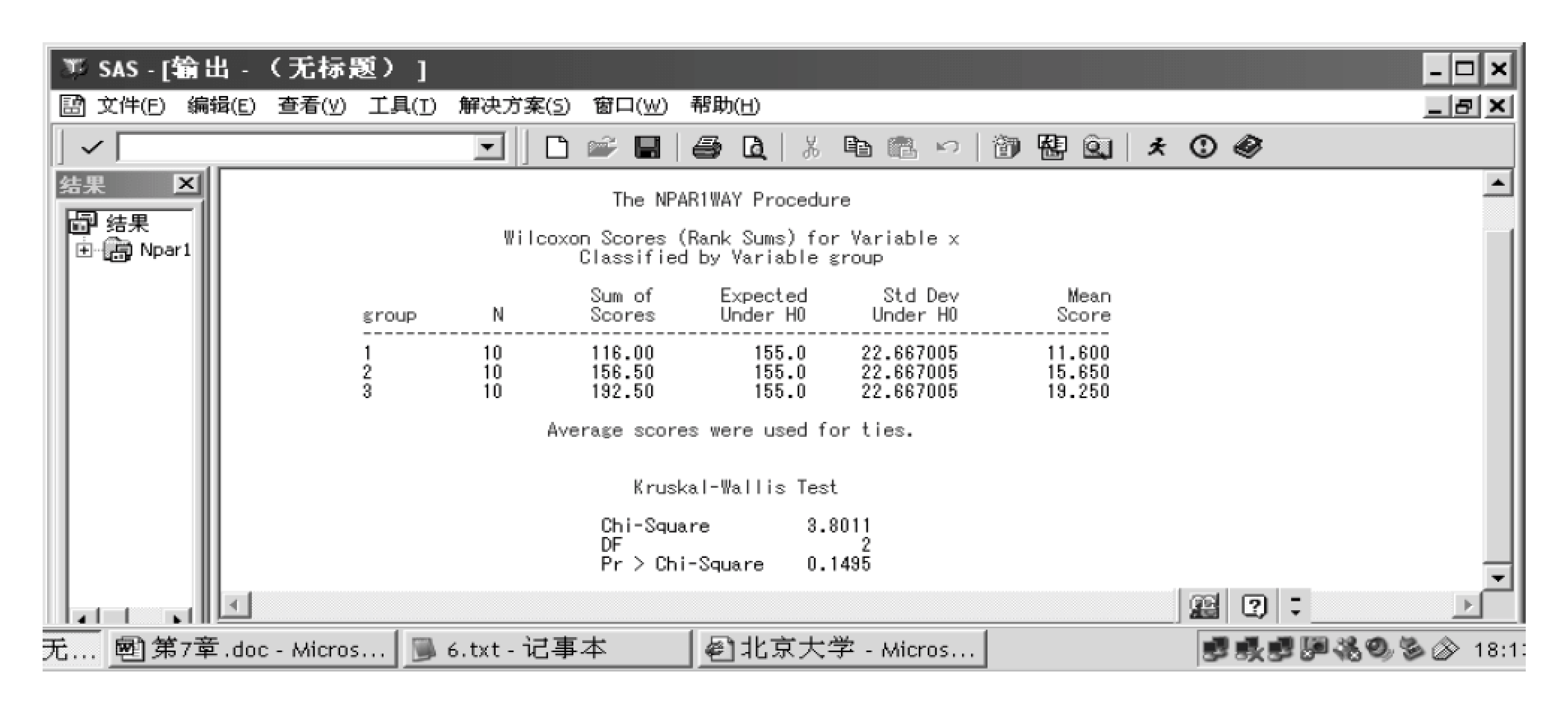
从图 7.5 看，WILCOXON 统计量 Z 为－1.1053。t 检验的概率为 0.2828 大于 α 值 0.05，所以没有足够的理由拒绝 H0，表明实验组与对照组消肿效果没有显著差异。

例 6：三样本(即三水平)均差的显著性检验，见程序 7.6。

程序 7.6：实验组与两个对照组消肿效果没有差异。

```
DATA dbs33;
INPUT group x @@;              /* group=1为对照组,其余为实验组 */
CARDS;
1 48 1 70 1 60 1 50 1 65 1 68 1 75 1 64 1 66 1 58
2 60 2 80 2 75 2 48 2 70 2 60 2 70 2 60 2 70 2 72
3 55 3 76 3 67 3 61 3 66 3 64 3 77 3 82 3 80 3 79
;
PROC  NPAR1WAY WILCOXON;
    CLASS group;               /* 指定 group 为分类变量 */
    VAR x;                     /* 指定数字型因变量 x */
```

运行程序 7.6 产生图 7.6 所示的结果。


The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable x
Classified by Variable group

group	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
1	10	116.00	155.0	22.667005	11.600
2	10	156.50	155.0	22.667005	15.650
3	10	192.50	155.0	22.667005	19.250

Average scores were used for ties.

Kruskal-Wallis Test

Chi-Square	3.8011
DF	2
Pr > Chi-Square	0.1495

图 7.6　三样本均差的显著性检验

结果分析：

H0：实验组与对照组消肿效果没有差异。

图 7.5 是双样本均差的显著性检验，图 7.6 是三样本均差的显著性检验。相比之下，两个图形的结果很不一样。双样本均差的显著性检验看图 7.5 中的 Z 检验，三样本显著性检验看图 7.6 中 Kruskal-Wallis 值。

从图 7.6 看，Kruskal-Wallis 值为 3.801，有(3－1＝2)个自由度，其卡方检验的概率为 0.1495 大于 α 值 0.05，所以没有足够的理由拒绝 H0，表明实验组与其他两个对照组消肿效果没有显著差异。

习　题　7

1. 简述两个均值的比较方法。

2. 试用 MEANS 过程及其 t 统计量对变量血糖与标准均值 5 进行两个均值差异的检验。

3. 配对样本 t 检验。对 20 位肿瘤患者，其中的 10 个人采用药物＋化疗治疗，另 10

个人采用药物＋放疗治疗。两周后测得体重增加见程序 A.33 中的数据处理(单位：公斤)。试对程序 A.33 中的数据做两种平均疗效的差异性检验。

程序 A.33：

```
DATA TZ;
INPUT v1 v2;
  D=v1-v2;                          /* 20位肿瘤患者体重之差 */
LABEL v1='药物+化疗' v2='药物+放疗';
CARDS;
0.90  0.75  0.80  0.81  0.69  0.78  0.53  0.66  0.51  0.59
0.80  0.85  0.98  0.82  0.66  0.63  0.60  0.51  0.67  0.68
PROC MEANS MEAN STD T PRT;          /* 计算最主要的统计量 */
    VAR  D;
RUN;
```

4. 将被试者分为对照组和实验组两组，使用同一种抗癌药物，一个月后测得肿块大小见程序 A.34 中的数据所示。试用“TTEST 过程及其 t 检验”做两组疗效差异性的检验。

程序 A.34：

```
DATA dbs;
INPUT group x @@;
CARDS;
1 78 1 90 1 90 1 50 1 85 1 88 1 85 1 88 1 84 1 88
2 40 2 80 2 75 2 48 2 70 2 60 2 70 2 60 2 70 2 62
;
PROC TTEST;
CLASS group;
VAR x;
RUN;
```

5. 用 WILCOXON 秩和检验对程序 A.35 中的实验组与对照组此双样本(即二水平)均差进行显著性检验。

程序 A.35：

```
DATA dbs;
INPUT group x @@;
CARDS;
1 78 1 90 1 90 1 50 1 85 1 88 1 85 1 88 1 84 1 88
2 40 2 80 2 75 2 48 2 70 2 60 2 70 2 60 2 70 2 62
;
PROC NPAR1WAY WILCOXON;
    CLASS group;          /* 指定 group 为分类变量 */
    VAR x;                /* 指定数字型因变量 x */
RUN;
```

第 8 章 chapter 8

方差分析

方差分析的基本原理是将全部数据的总均值的离差平方和分解为若干部分，每一部分表示某因素的主效应、交互效应所产生的影响。然后将各部分(如组间平方和、误差平方和)的均方与误差均方进行比较，进而确定哪些因素或哪些交互作用比较重要和最重要。

方差分析的作用是观察实验的结果究竟受到哪些因素(自变量)、哪些水平的影响。

方差分析的公式为：总变异性＝组内变异＋组间变异。

常用方差分析法有下列 4 种：

- 完全随机设计数据的方差分析(即单因素方差分析)
- 随机区组数据的方差分析(即二因素方差分析)
- 拉丁方设计数据的方差分析(即三因素方差分析)
- R * C 析因设计数据方差分析(有交互作用的方差分析)

方差分析的过程命令：

(1) PROC ANOVA；

用于每个因素水平(每组)的个案数目相等的情形，即均衡数据的方差分析。如果只考虑组内变异和组间变异(One-Way 单向方差分析)时，ANOVA 也能处理非均衡数据。

(2) PROC GLM；

用于每个因素水平(每组)的个案数目不相等的情形，即非均衡数据的方差分析。

8.1 用 ANOVA 做均衡数据的方差分析

均衡数据指因素每个水平(如男女)的个案数目相等。

8.1.1 ANOVA 过程命令

1. 命令格式

```
PROC ANOVA;
    CLASS v1;
    MODEL y= v1 v2;
```

```
[MEANS v3 选项;]
[ALPA=P]      /* 默认为显著性水平 P=0.05* /
```

2. 选项说明

[MEANS v3 选项;]和[ALPA=P]语句是任选项。[MEANS v3 选项;]用于产生各个效应项的因变量的均值。若指定了此选项,则检验主效应间的均值。

[MEANS v3 选项;]中的选项如下(任选 1—2 项):

BON、DUNCAN、LSD、REGWF、REGWQ、SNK(Q 检验。常用)、SCHEFFE、SIDAK、SMM(GT2)、TUKEY、WALLER、DUNNETT(常用)。

8.1.2 单因素方差分析

单因素方差分析也称为完全随机设计数据的方差分析。

例如某研究所对 3 种装修涂料(a、b、c)进行甲醛含量检测,每种涂料各做 5 次检测,所得的数据见表 8.1。试检验各种涂料甲醛平均含量有无显著差异。

表 8.1 各种涂料甲醛含量

涂料 a	涂料 b	涂料 c	涂料 a	涂料 b	涂料 c
3.5	3.8	4.0	3.0	3.1	3.3
3.4	3.7	3.9	3.2	3.6	3.8
3.9	3.4	3.6			

根据题意编辑出程序 8.1。

程序 8.1:

```
DATA f1;
INPUT t1 x @@;          /* 定义涂料变量 t1 和甲醛含量变量 x* /
LABEL t1='涂料' x='甲醛含量';
CARDS;
1 3.5 2 3.8 3 4.0 1 3.4 2 3.7 3 3.9
1 3.9 2 3.4 3 3.6 1 3.0 2 3.1 3 3.3
1 3.2 2 3.6 3 3.8
;
PROC FORMAT;
VALUE t1F 1='涂料 a' 2='涂料 b' 3='涂料 c';
FORMAT t1 t1F.;         /* 将编码值赋予原变量 t1,即确认 * /
PROC ANOVA;
    CLASS t1;           /* 定义 t1 为分类变量 * /
    MODEL x=t1;         /* x 指定 x 为数字型因变量 * /
```

运行程序 8.1 产生图 8.1 所示的结果。

结果分析:

H0:各种涂料甲醛平均含量没有显著差异。

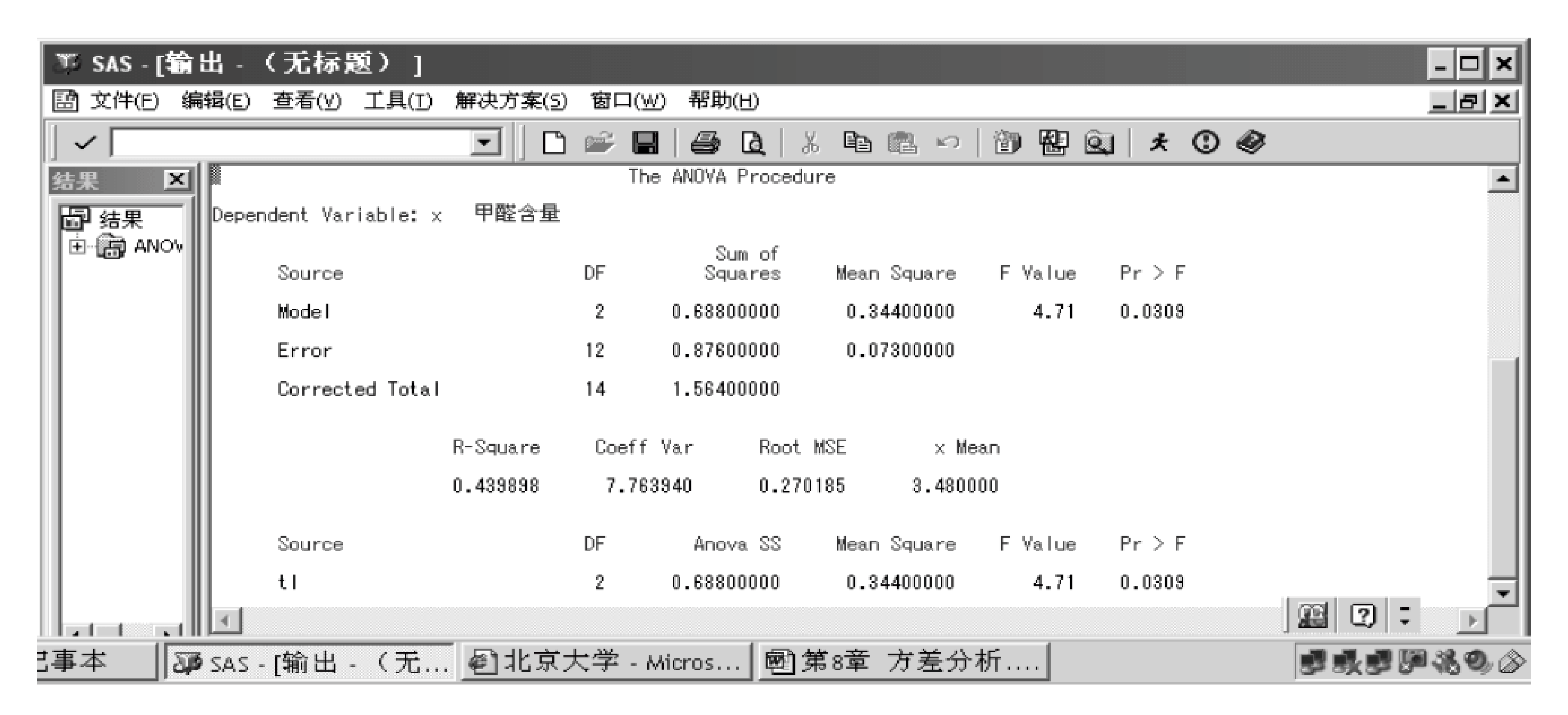


图 8.1　各种涂料甲醛平均含量的显著差异检验

检验：

从图 8.1 看，模型的显著性水平 0.0309<α 值 0.05，显著。因此可以继续观察各个因素的差异。

从图 8.1 的因素 tl 一行看，显著性水平 0.0309<α 值 0.05，显著。所以有足够的理由拒绝 H0，说明 3 组涂料甲醛平均含量有显著差异。

再看图 8.1 的 R-Square（单向方差分析）值为 0.439898，说明总体方差只有 44%是来自组间变异，不太理想。

8.1.3　双因素方差分析

双因素方差分析也称为随机区组数据的方差分析。

例如对血小板偏低者用 4 种不同的药物治疗后血小板的数据见表 8.2。试检验 4 种药物平均疗效有无显著差异。

表 8.2　用 4 种不同的药物治疗血小板偏低者的数据

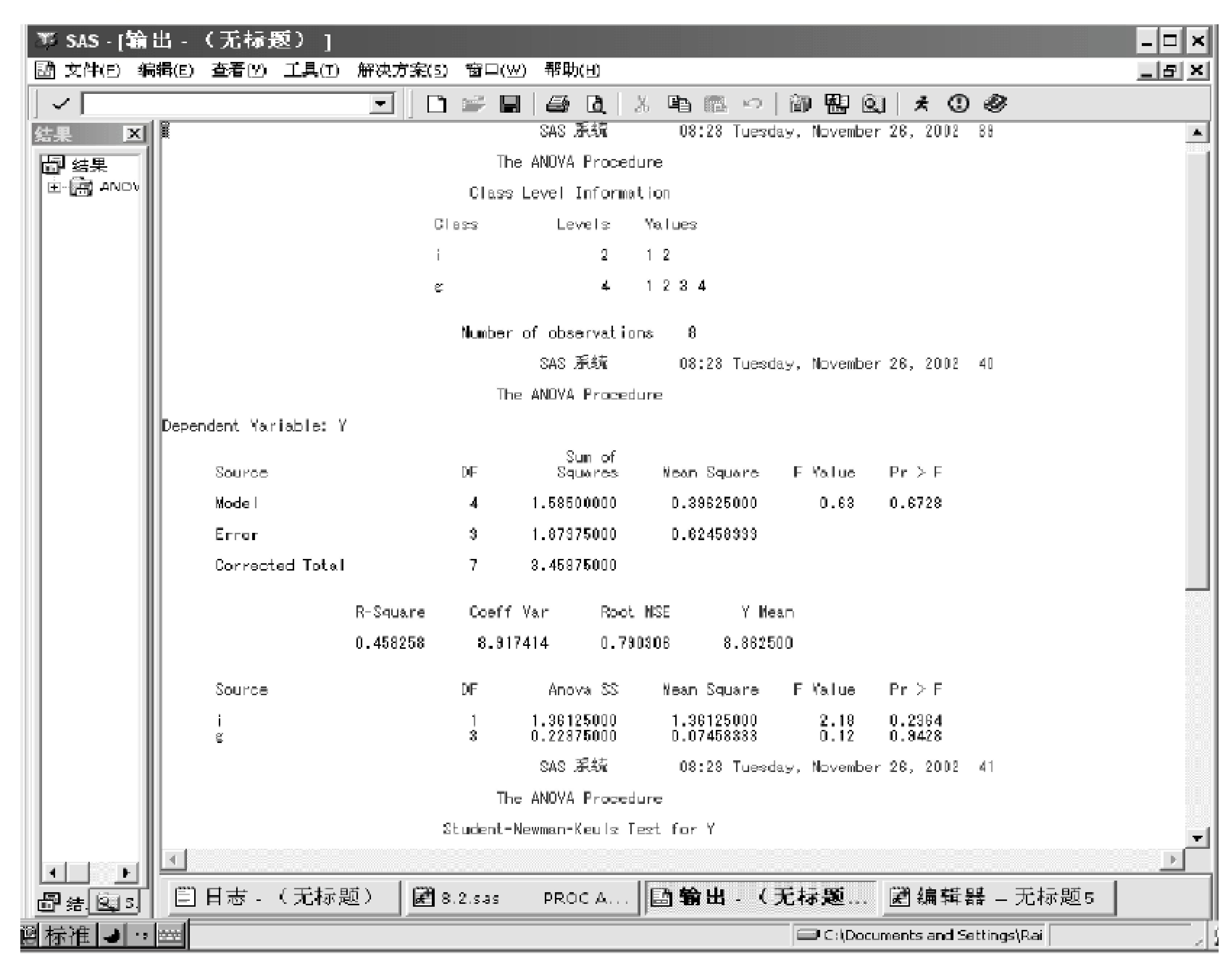
区间（受试者）/编号	处理组（4 种疗效）			
	1	2	3	4
1	9.1	9.2	9.6	10.1
2	8.5	9.0	9.2	9.5
3	8.0	8.6	9.0	9.2
4	8.2	8.8	8.9	10.5
5	8.5	9.0	9.1	9.5
6	9.0	9.2	9.4	9.6
7	9.5	9.8	10.1	10.5
8	10.1	10.3	10.5	10.8

根据题意编辑出程序 8.2。

程序 8.2：

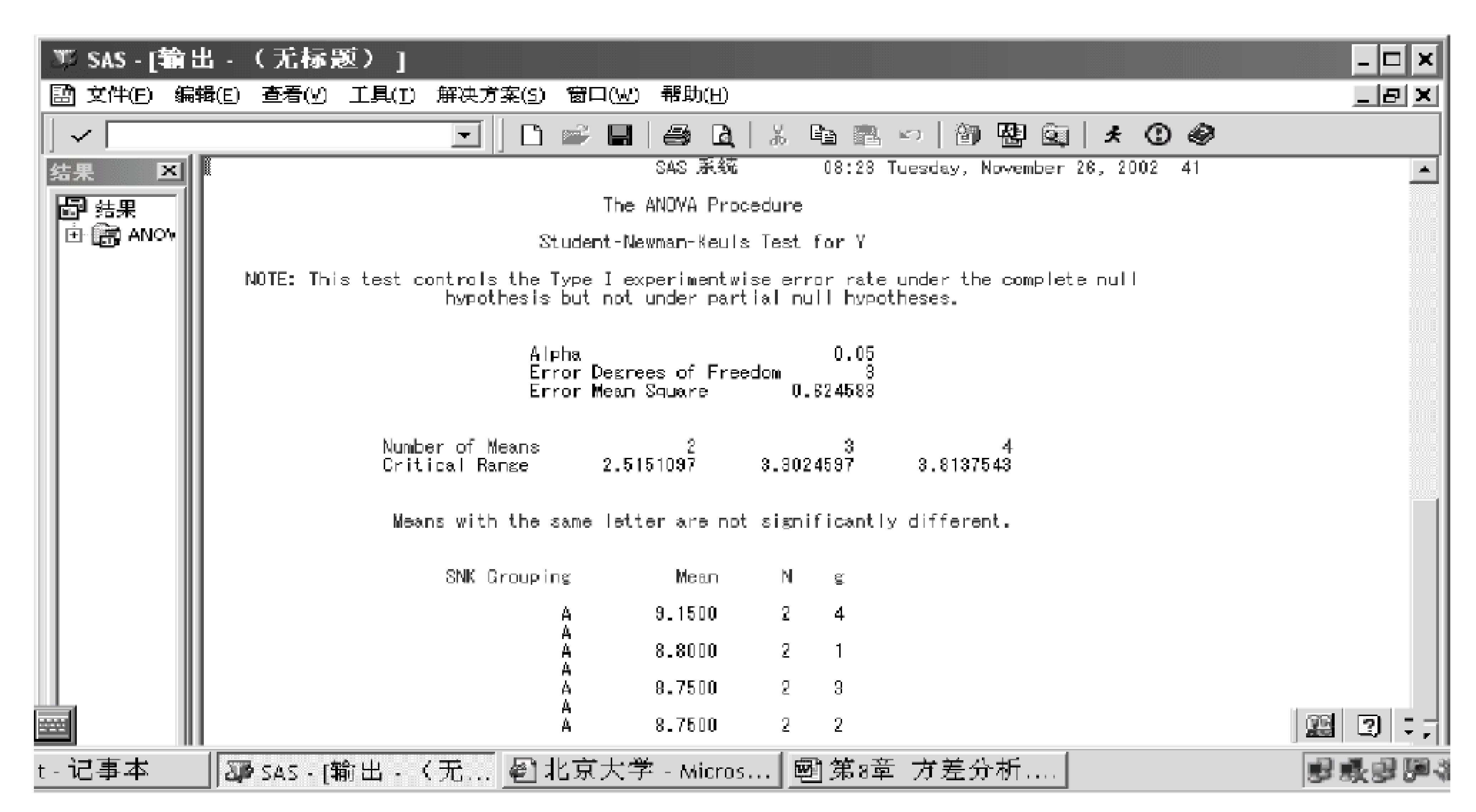
```
DATA f2;
     DO i=1 TO 8;
        DO g=1 TO 4;
         INPUT Y @@;
         OUTPUT;
        END;
      END;
CARDS;
9.1    9.2    9.6    10.1
8.5    9.0    9.2    9.5
8.0    8.6    9.0    9.2
8.2    8.8    8.9    10.5
8.5    9.0    9.1    9.5
9.0    9.2    9.4    9.6
9.5    9.8    10.1   10.5
10.1   10.3   10.5   10.8
;
PROC ANOVA;
   CLASS i g;
   MODEL Y=i g;
   MEAN g/SNK;      /* 增加 SNK 两两比较的功能 */
```

运行程序 8.2 产生图 8.2 所示的结果。



(a) 4种药物平均疗效有无显著差异的检验 1

图 8.2 疗效显著差异检验



(b) 4种药物平均疗效有无显著差异的检验 2

图 8.2（续）

结果分析：

H0：治疗后 4 组血小板平均含量没有显著差异。

检验：

从图 8.2(a)看，模型的显著性水平 0.6728>α 值 0.05，非常不显著。模型不佳。

从图 8.2(a)的因素 i 一行看，显著性水平 0.2364>α 值 0.05，不显著。所以没有足够的理由拒绝 H0，说明 4 组血小板平均含量没有显著差异。

同理，8 位受试者血小板平均含量没有显著差异。

再看图 8.1 的 R-Square（单向方差分析）值为 0.439898，说明总体方差只有约 44% 是来自组间变异，不理想。

图 8.2(b)显示组与组之间疗效的检验，由于 4 组的 SNK 检验结果都显示 A 字母，表明 4 组的疗效没有差异。

结论：应该更换数据或模型。

8.1.4 三因素方差分析

三因素方差分析也称为拉丁方设计数据的方差分析。对于三因素方差分析，假定各因素间的交互作用可忽略不计或不存在交互作用，而且各因素的水平都有相同。见表 8.3 的数据，有 4 位患者分别服用 4 种药物，4 个月内测得体内总胆红素（单位：μ mol/L），要求检验 4 组药物作用于每位患者后体内平均总胆红素有无显著差异。

表 8.3　4 组药物作用于每位患者后体内平均总胆红素(单位：μ mol/L)

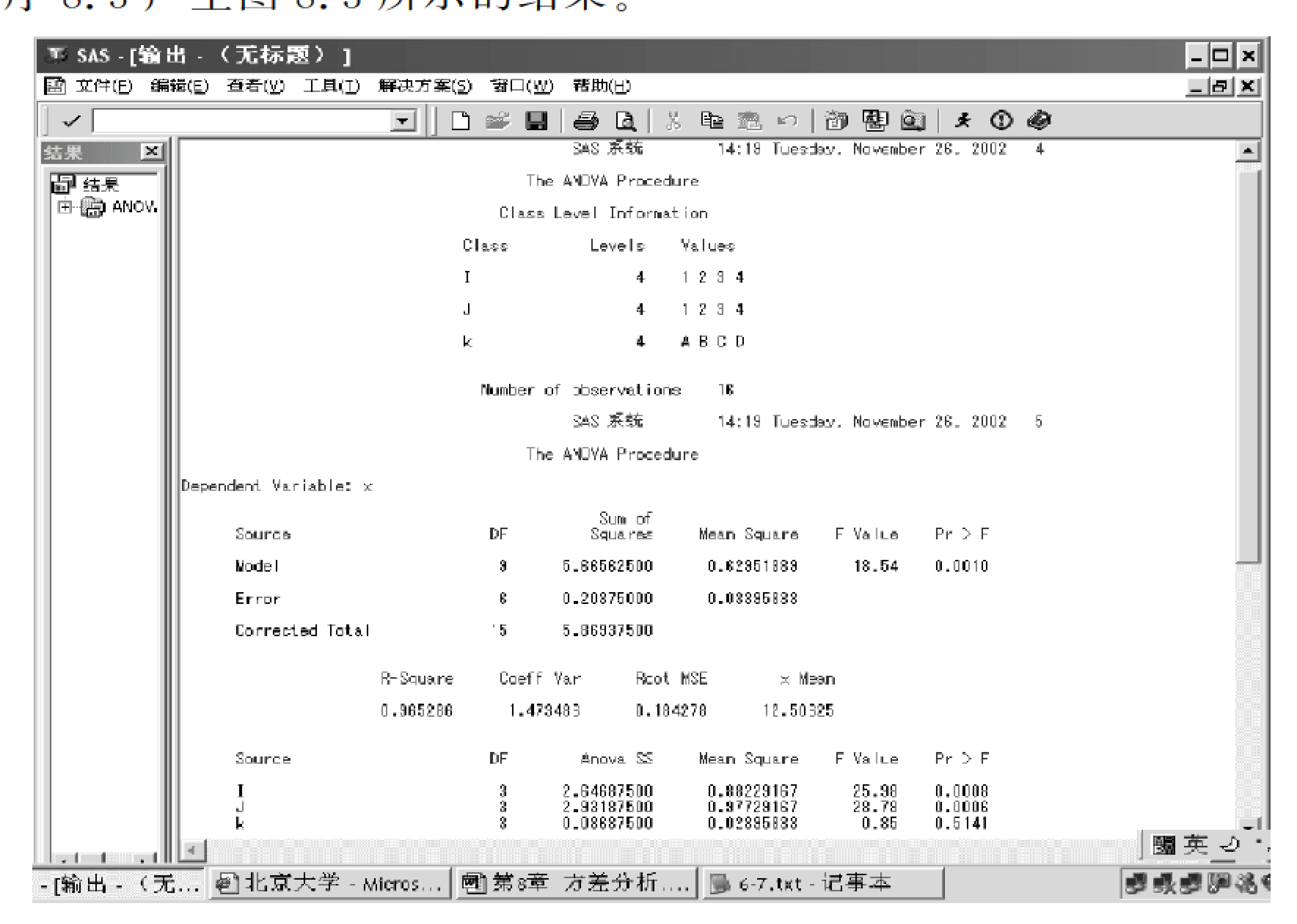
试验月份	4 位受试者			
	甲	乙	丙	丁
1	A11.5	B11.8	C12.3	D12.5
2	B11.8	C12.1	D12.5	A12.8
3	C12.0	D12.5	A12.8	B13.0
4	D12.5	A12.8	B13.1	C14.1

解法：见程序 8.3。

程序 8.3：用 DUNNETT 检验。

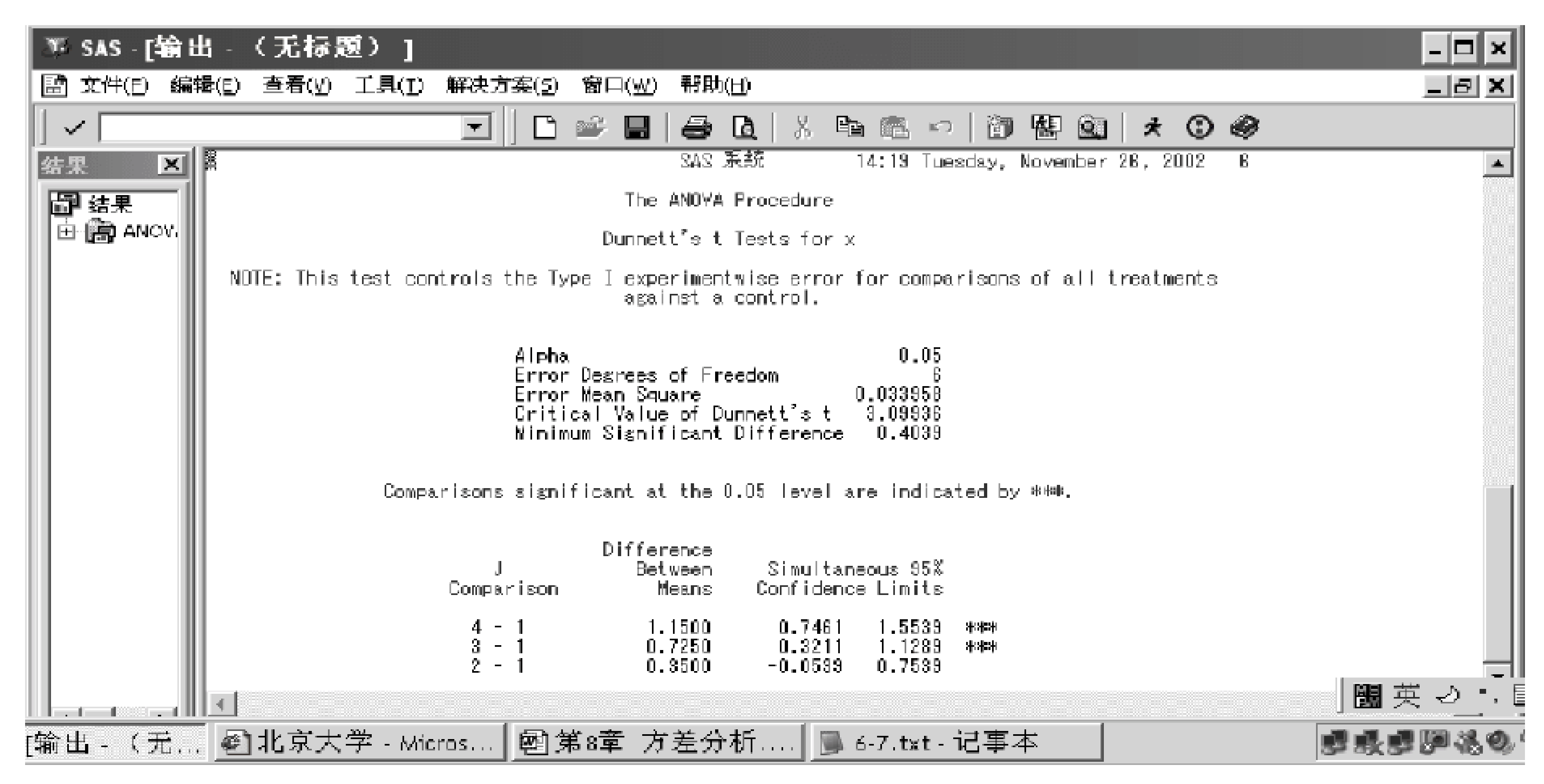
```
DATA f3;
  DO I=1 TO 4;                /* I为试验的月份 */
    DO J=1 TO 4;              /* J为 4位受试者 */
     INPUT k $  x @@ ;        /* K为 4种药物,x为总胆红素 */
     OUTPUT;
    END;
  END;
CARDS;
A 11.5 B 11.8 C 12.3 D 12.5 B 11.8 C 12.1 D 12.5 A 12.8
C 12.0 D 12.5 A 12.8 B 13.0 D 12.5 A 12.8 B 13.1 C 14.1
PROC ANOVA;
  CLASS i j k;                /*  i j k为分组变量 */
  MODEL x=i j k;              /*  x为总胆红素 (单位:μ mol/L) */
  MEANS j/DUNNETT('1');       /* 增加 DUNNETT比较功能,('1')表示 j=1为对照组 */
```

运行程序 8.3 产生图 8.3 所示的结果。



(a) 模型检验

图 8.3　体内平均总胆红素差异性检验



(b) 组与组的均值比较

图 8.3（续）

结果分析：

H0：体内平均总胆红素没有显著差异(模型不佳)。

检验：

从图 8.3(a)看，模型的显著性水平 0.0010<α 值 0.05，非常显著。说明体内平均总胆红素有显著差异。

从图 8.3(a)的因素 i 和 j 行看，显著性水平都小于 α 值 0.05，显著。所以试验的月份不同，体内平均总胆红素就有显著差异。不同的受试者体内平均总胆红素也有显著差异。但是从图 8.3(a)的因素 k 行看，4 种药物的疗效没有显著差异。

再看图 8.3(a)的 R-Square(单向方差分析)值为 0.965286，说明总体方差中有 96.5%是来自组间变异，很理想。

图 8.3(b)显示受试者第 2 组与对照组(第 1 组)之间的平均总胆红素没有显著差异。而受试者第 4 组与对照组(第 1 组)之间的平均总胆红素有显著差异(有 * * * 记号)，受试第 3 组与对照组(第 1 组)之间的平均总胆红素也有显著差异(有 * * * 记号)。

8.1.5 R* C交互因素的方差分析

R * C 交互因素的方差分析又称为 R * C 析因设计数据方差分析，它是有交互作用的方差分析。

例如某医院对 8 例缺钙患者，分为 4 组，进行不同药物治疗，两个月后检测血中含钙量(单位：μ 克/100ml)如表 8.4 所示。

解法：根据题意编辑出程序 8.4a。

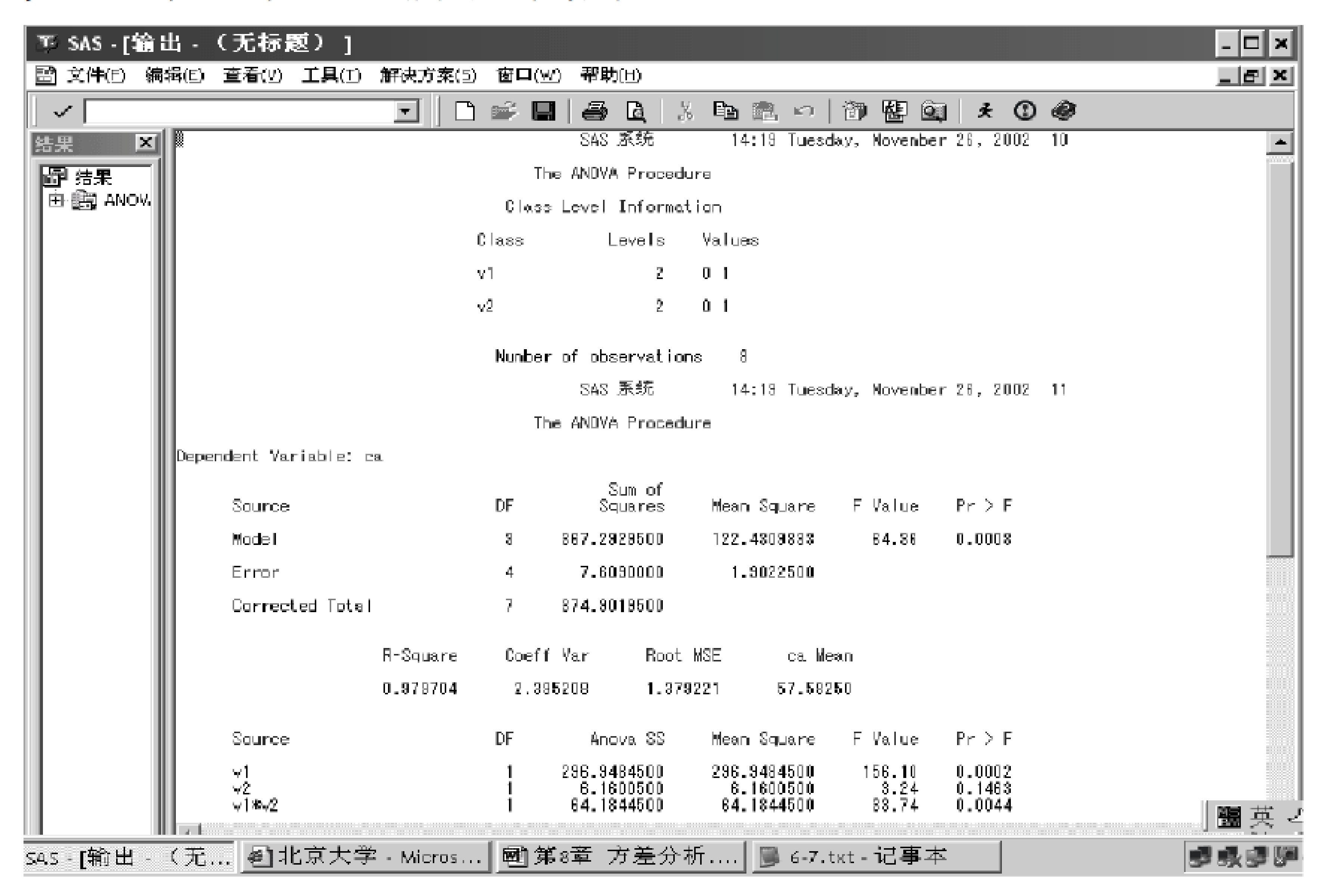
程序 8.4a：

表 8.4　4 组缺钙患者血中含钙量(单位：μ 克/100ml)

	甲药(v1)			
	不用		用	
乙药(v2) 不用	54.6	55.8	60.68	62.76
乙药(v2) 用	48.8	46.76	64.48	66.78

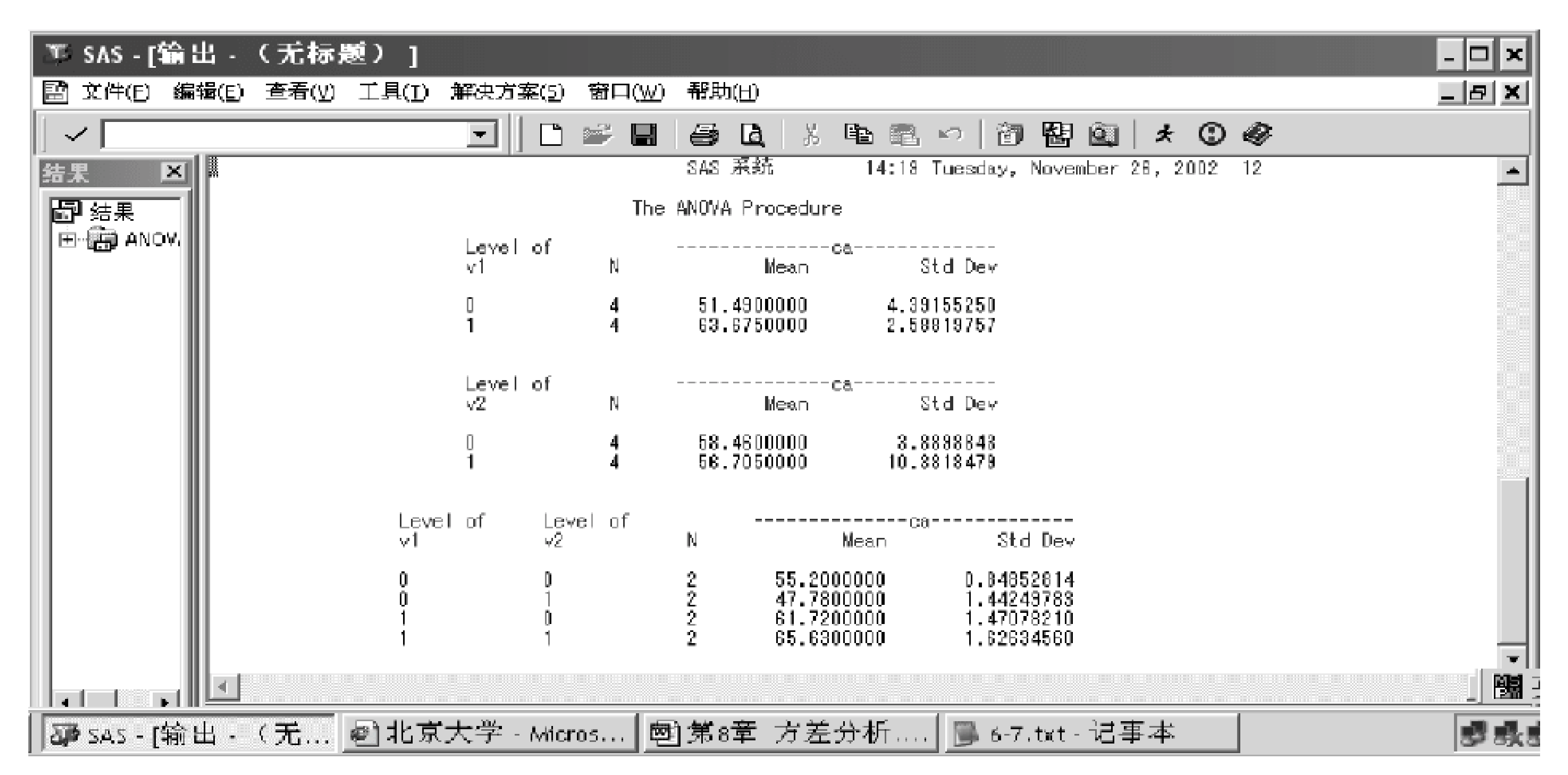
```
DATA f4;
INPUT id v1 v2 ca;          /* v1为甲药,v2为乙药 */
CARDS;
01 0 0 54.60
02 0 0 55.80
03 1 0 60.68
04 1 0 62.76
05 0 1 48.80
06 0 1 46.76
07 1 1 64.48
08 1 1 66.78
;
PROC ANOVA;
  CLASS v1 v2;
  MODEL ca=v1 v2 v1*v2;
MEANS v1 v2 v1*v2;          /* 进一步比较各因素中补钙的均值差异 */
```

运行程序 8.4a 产生图 8.4 所示的结果。



(a) 模型检验

图 8.4　4 组缺钙患者血中含钙量差异性检验



(b) 均值比较

图 8.4（续）

结果分析：

H0：4 组缺钙患者血中平均含钙量没有差异(模型不佳)。

检验：

从图 8.4(a)看，模型的显著性水平 0.0008<α 值 0.05，非常显著。说明治疗后患者体内平均含钙量有显著差异。

从图 8.4(a)的因素 v1 行看，显著性水平 0.0002 小于 α 值 0.05，显著。所以甲药用与不用，对患者平均含钙量有显著差异。

从图 8.4(a)的因素 v1 * v2 行看，显著性水平 0.0044 小于 α 值 0.05，显著。所以甲药与乙药有显著的交互作用，对患者平均含钙量有显著影响。

但是从图 8.4(a)的因素 v2 行看，显著性水平 0.1463 大于 α 值 0.05，很不显著。所以乙药用与不用，对患者平均含钙量没有显著差异。

程序 8.4b：将程序 8.4a 改为用 DO…END 语句输入数据。

```
DATA f4b;
  DO v2=0 TO 1;            /* v1 为甲药,v2 为乙药 */
    DO v1=0 TO 1;
     DO I=1 TO 2;
       INPUT ca @@;
       OUTPUT;
     END;
    END;
  END;
CARDS;
54.60 55.80 60.68 62.76 48.80 46.76 64.48 66.78
;
PROC ANOVA;
```

```
  CLASS v1 v2;
  MODEL ca= v1 v2 v1 * v2;
MEANS v1 v2 v1 * v2;          /* 进一步比较各因素中补钙的均值差异 * /
```

运行程序 8.4b 也产生图 8.4 所示的结果。

比较程序 8.4a 和程序 8.4b 所产生的输出,其结果一模一样。请读者亲自上机体会。

8.1.6 多个实验组与对照组的均值比较

已知某个因素有 3 个水平,这时可有 3－1、2－1、3－2 一共 3 对均值的比较。如果这时将第 1 组作为对照组,则有"3－1"和"2－1"一共两对均值的比较,这就是多个实验组与对照组的均值比较。

解法:已知实验组与对照组的数据如表 8.5 所示,采用 DUNNETT 检验法,见程序 8.5。

程序 8.5:

```
DATA f5;
INPUT group weight @ @ ;
CARDS;
1 0.50 1 0.75 1 0.80 1 0.91 1 0.69 1 0.48 1 0.33 1 0.66 1 0.51 1 0.59
2 0.60 2 0.68 2 0.78 2 0.82 2 0.56 2 0.23 2 0.40 2 0.51 2 0.47 2 0.48
3 0.80 3 0.75 3 0.80 3 0.85 3 0.68 3 0.50 3 0.35 3 0.62 3 0.50 3 0.55
;
PROC ANOVA;
  CLASS group;
  MODEL weight = group;
MEANS group/DUNNETT('3');      /* 第 3组为对照组 * /
```

表 8.5 实验组与对照组平均体重增加的数据

治 疗 方 法	1	2	3	4	5	6	7	8	9	10
V1:药物＋化疗	0.50	0.75	0.80	0.91	0.69	0.48	0.33	0.66	0.51	0.59
V2:药物＋放疗	0.60	0.68	0.78	0.82	0.56	0.23	0.40	0.51	0.47	0.48
V3: 药物＋放化疗	0.80	0.75	0.80	0.85	0.68	0.50	0.35	0.62	0.50	0.55

运行程序 8.5 也可产生图 8.5 所示的结果。

结果分析:

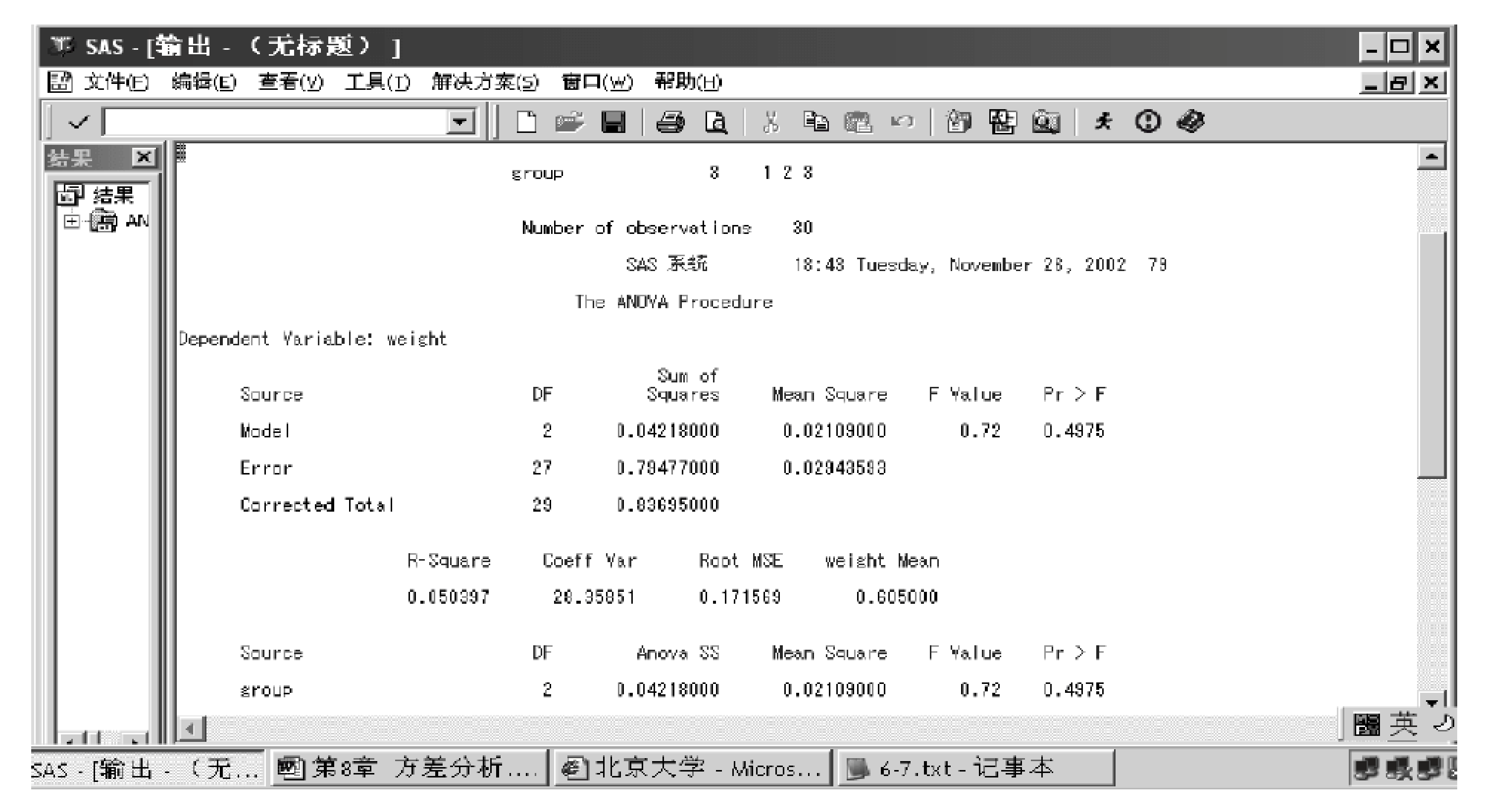
H0:实验组与对照组平均体重平均增加没有差异(模型不佳)。

检验:

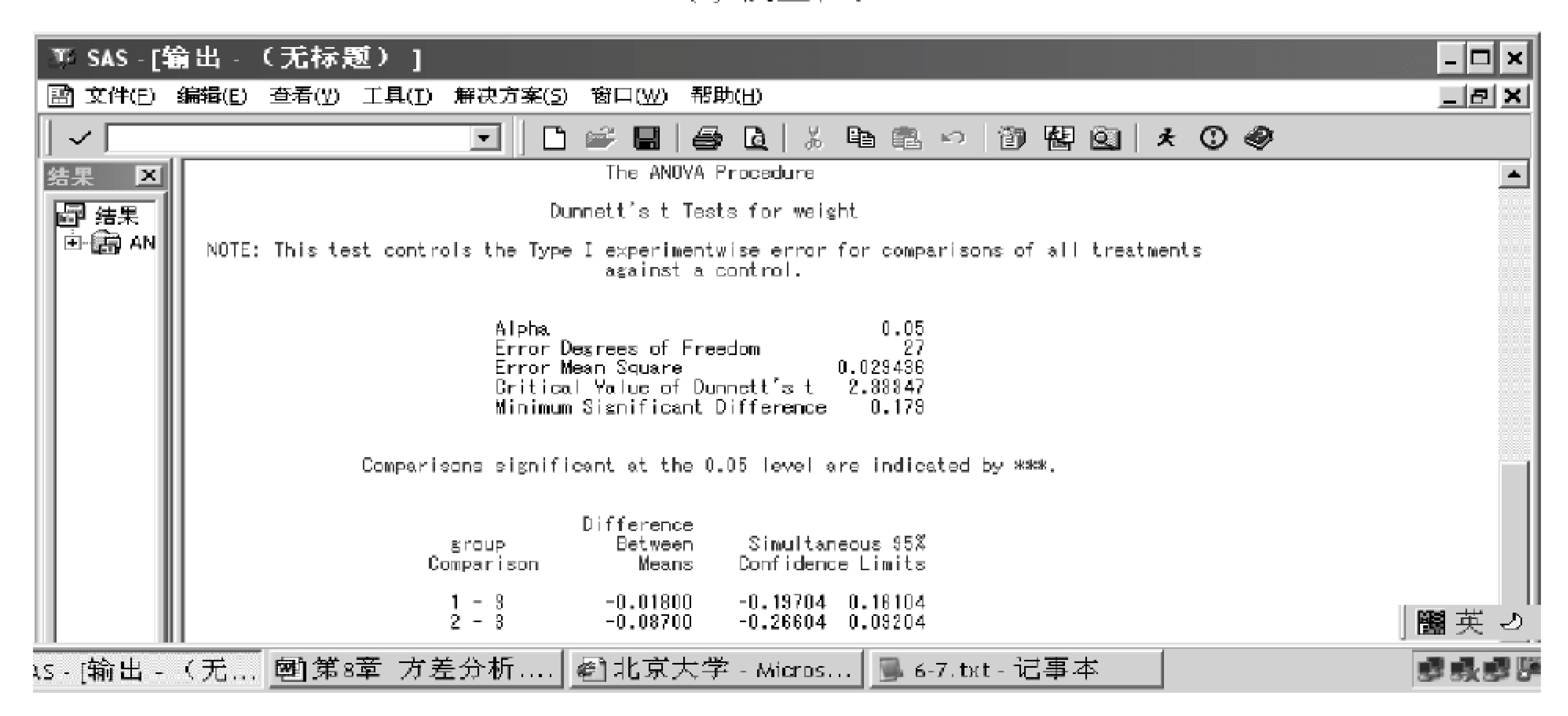
从图 8.5(a)看,模型的显著性水平 0.4975>α 值 0.05,非常不显著。说明实验组与各对照组平均体重增加没有差异。

从图 8.5(a)的 group 因素看,显著性水平 0.4975>α 值 0.05,非常不显著。所以各对照组之间平均体重增加没有差异。

从图 8.5(b)的组与组均值差异性检验看,行尾均无"* * *",表明平均体重增加没



(a) 模型检验



(b) 组与组均值差异性检验

图 8.5　实验组与对照组平均体重增加的差异性检验

有差异。

8.1.7　用 SNK 的 Q 检验法比较组间均值

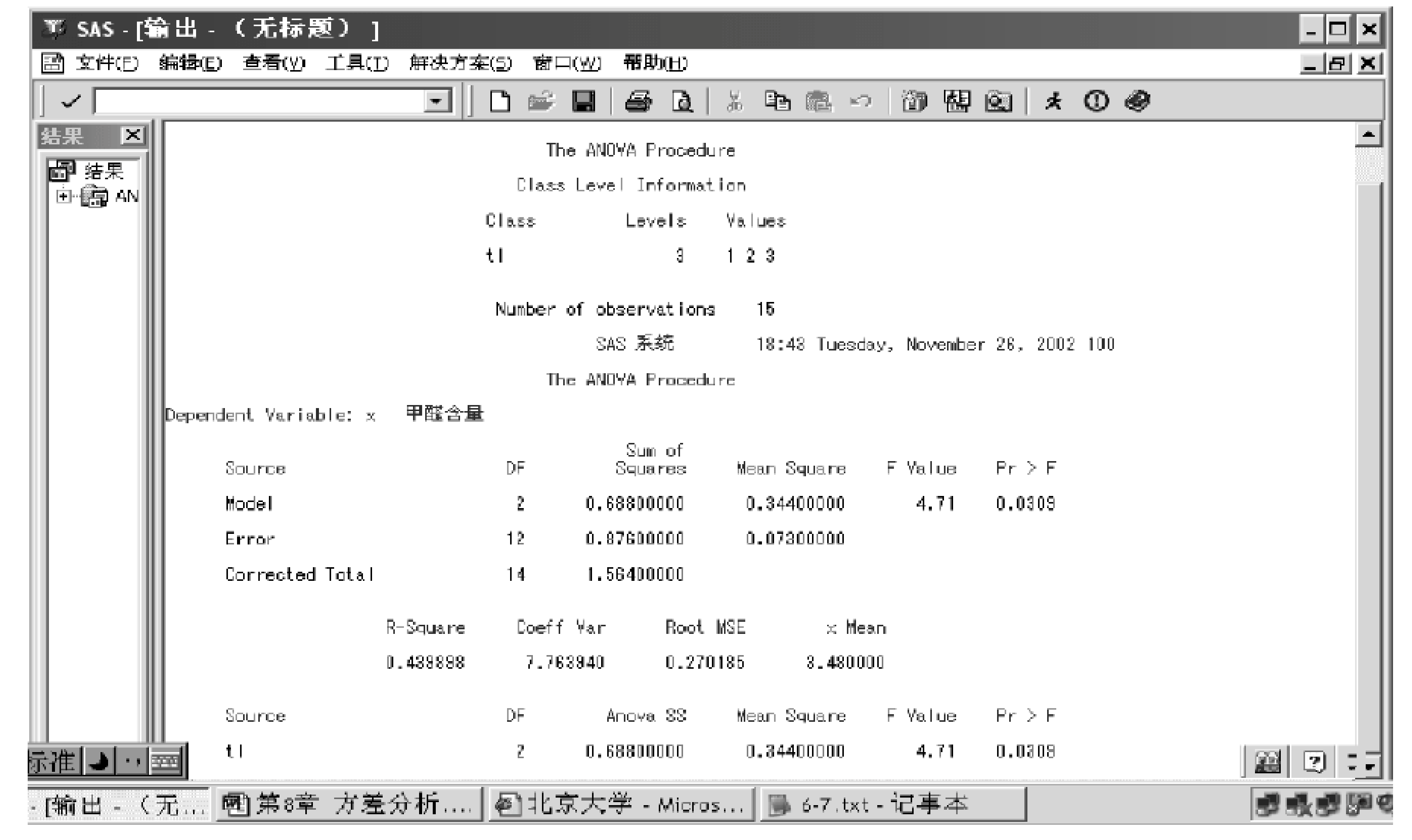
SNK 的 Q 检验法可用于单因素 3 水平(或 3 水平以上)的组间均值比较。对程序 8.1 中的 3 种涂料数据(见表 8.1),用 SNK 的 Q 检验法检验各组平均甲醛含量的差异性。

程序 8.6:对程序 8.1 用 SNK 的 Q 检验法检验各组平均甲醛含量的差异性。

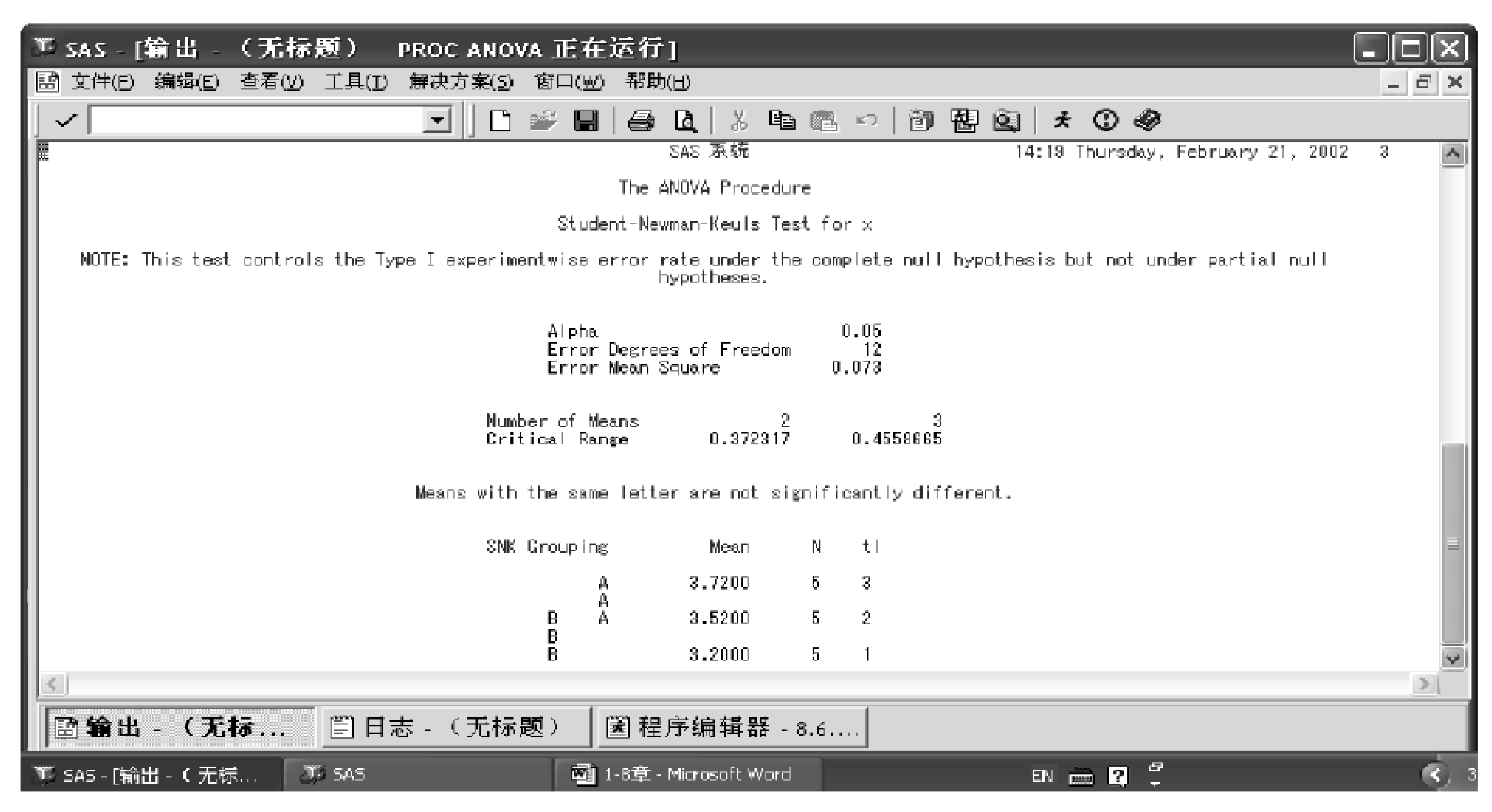
```
DATA f1;
INPUT t1 x @@;              /* 定义涂料变量 t1 和甲醛含量变量 x* /
LABEL t1='涂料' x='甲醛含量';
CARDS;
1 3.5 2 3.8 3 4.0 1 3.4 2 3.7 3 3.9
```

```
1 2.9 2 3.4 3 3.6 1 3.0 2 3.1 3 3.3
1 3.2 2 3.6 3 3.8
;
PROC FORMAT;
VALUE tlF 1= '涂料 a' 2= '涂料 b' 3= '涂料 c';
FORMAT tl tlF.;             /* 将编码值赋予原变量 tl,即确认 */
PROC ANOVA;
    CLASS tl;               /* 定义 tl 为分类变量 */
    MODEL x= tl;            /* 指定 x 为数字型因变量 */
MEANS tl/SNK;               /* 用 SNK 的 Q 检验法检验 */
```

运行程序 8.6 产生图 8.6 所示的结果。



(a) 模型检验



(b) 两组两两比较

图 8.6　用 SNK 的 Q 检验法检验各组平均甲醛含量的差异性

结果分析：

H0：3 组涂料平均甲醛含量没有显著差异。

检验：

从图 8.6(a)看，模型的显著性水平 0.0309<α 值 0.05，非常显著。模型好。

从图 8.6(a)的因素 tl 一行看，显著性水平 0.0309<α 值 0.05，非常显著。所以有足够的理由拒绝 H0，说明 3 组涂料平均甲醛含量有显著差异。

从图 8.6(b)进一步看到，第 2 组涂料(TL)和第 3 组涂料平均甲醛含量有显著差异(用字母 A 和 B 表示)，第 1 组涂料(TL)和第 2 组涂料平均甲醛含量有显著差异(用字母 A 和 B 表示)，第 1 组涂料(TL)和第 3 组涂料平均甲醛含量有显著差异(用字母 A 和 B 表示)。

8.2 用 GLM 进行非均衡数据方差分析

非均衡数据指因素的各个水平的个案数目不等，例如性别中男女人数可以不等。GLM 是(General Linear Model)的缩写。

8.2.1 GLM过程命令

1. GLM 过程命令格式

```
PROC GLM;
    CLASS v1 v2;
    MODEL 因变量 Y=v1 v2 v1 * v2;
    MEANS v3 选项;
```

2. 选项说明：参阅 8.1.1 节

[MEANS v3 选项;]用于产生各个效应项的因变量均值。若指定了此选项，则检验主效应间的均值。

[MEANS v3 选项;]中的选项如下(任选 1—2 项)：

BON、DUNCAN、LSD、REGWF、REGWQ、SNK(Q 检验。常用)、SCHEFFE、SIDAK、SMM(GT2)、TUKEY、WALLER、DUNNETT(常用)。

8.2.2 GLM过程的统计功能

1. GLM 胜任以下的统计分析

- 一元回归(简单回归)
- 多元回归及多重回归
- 方差分析(对非均衡数据更佳)
- 协方差分析

- 反应面模型分析
- 加权回归
- 多项式回归
- 偏相关分析
- 多元方差分析
- 重复测量方差分析

2. GLM 过程中的建模形式(见表 8.6)

表 8.6 GLM 的建模形式

模型说明	模型形式	模型说明	模型形式
MODEL Y=a b c;	主效应	MODEL Y1 Y2=a b;	多元方差分析
MODEL Y=a b a*b;	主效应+交互效应	MODEL Y=a x;	协方差分析
MODEL Y=a b a(b);	主效应+嵌套效应		

说明：表 8.6 中，a、b、c 为分类变量，x、y 是连续变量(次序以上类型的变量)。

8.2.3 用 GLM 做单因素 3 水平方差分析

单因素 3 水平指模型中有一个自变量，其值有 3 个水平(即 3 组)。因变量是数字型变量，例如人体内的白细胞数目、红细胞数目，或人的身高、体重等。

现有三组不同年龄的受试者，体检时测得体内的红细胞数目如表 8.7 所示。

表 8.7 体内的红细胞数目(单位：T/L)

40 岁以下	4.35	5.50	4.70	4.80	5.35	4.80	5.40	4.75	5.15
41～50 岁	4.65	3.50	5.30	4.60	5.15	4.75	5.10	3.85	
51 岁以上	4.60	5.20	5.10	4.50	4.95	5.15	4.95		

问：各组平均红细胞数目有无显著差异?

解法：见程序 8.7。

程序 8.7：

```
DATA RBC;
  DO J=1 TO 3;
  INPUT n;
    DO I=1 TO n;
     INPUT x @@;
     OUTPUT;
    END;
   END;
CARDS;
9
```

```
4.35 5.50 4.70 4.80 5.35 4.80 5.40 4.75 5.15
8
4.65 3.50 5.30 4.60 5.15 4.75 5.10 3.85
7
4.60 5.20 5.10 4.50 4.95 5.15 4.95
;
PROC GLM;
    CLASS j;        /* 定义 j 为分类变量 */
    MODEL x=j;      /* 指定 x 为数字型因变量 */
MEANS tl/SNK;       /* 用 SNK 的 Q 检验法检验 */
```

运行程序 8.7 产生图 8.7 所示的结果。

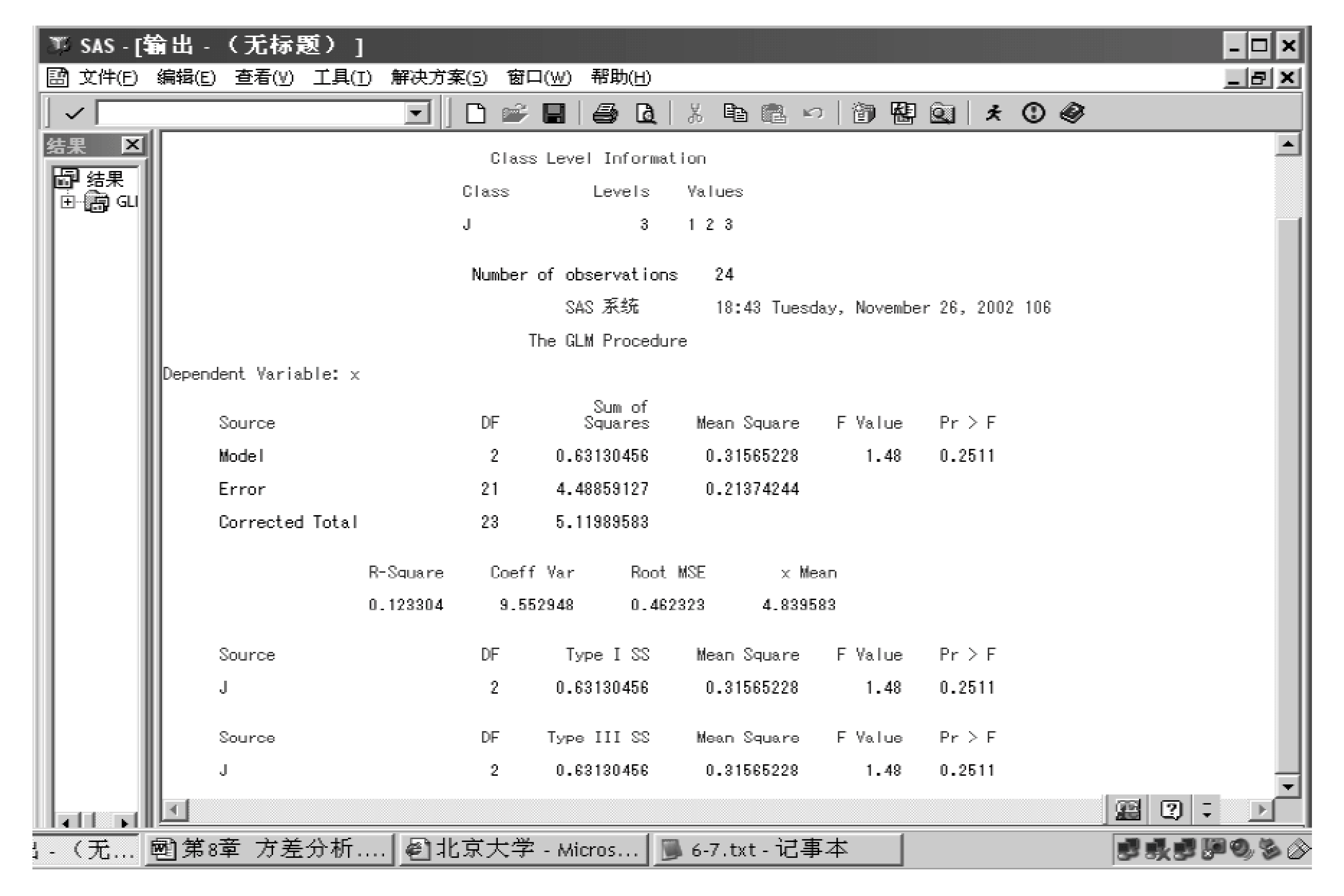


图 8.7　各组平均红细胞数目差异性的检验

H0：各组平均红细胞数目没有显著差异。

检验：

从图 8.7 看，模型的显著性水平 0.2511>α 值 0.05，非常不显著。模型不佳。

从图 8.7 的因素 j 一行看，显著性水平 0.2511>α 值 0.05，非常不显著。所以没有足够的理由拒绝 H0，说明各组平均红细胞数目没有显著差异。

8.2.4　用 GLM 做双因素方差分析

双因素方差分析指自变量有两个，因变量至少一个。现有男女两组受试者，分别服用三种减肥药物两个月，测得体重减轻（单位：kg）数据见表 8.8。问：三种减肥药物对男女两组受试者的减肥效果有无显著差异？

解法：见程序 8.8。

程序 8.8：

表 8.8 测得体重减轻数据(单位:kg)

		因素 B(药物)		
因素 A	A1(男)	B1	B2	B3
		2.5 2.7	2.3 3.1 2.8	4.1 3.9 3.7 4.3
	A2(女)	2.9 3.0	3.1 3.5 2.9	4.5 4.1 4.0 3.9
		3.5 3.6	2.8 3.2 3.6	4.7 4.3 4.5 5.0

```
DATA JF;
  DO a=1 TO 2;
    DO b=1 TO 3;
     INPUT n;
      DO I=1 TO n;              /* 每组重复读数的次数 */
       INPUT x @@;
       OUTPUT;
      END;
    END;
  END;
  DROP I n;
CARDS;
2
2.5 2.7
3
2.3 3.1 2.8
4
4.1 3.9 3.7 4.3
4
2.9 3.0 3.5 3.6
6
3.1 3.5 2.9 2.8 3.2 3.6
8
4.5 4.1 4.0 3.9 4.7 4.3 4.5 5.0
;
PROC GLM;
    CLASS a b;                  /* 定义 a b为分类变量 */
    MODEL x=a b a*b;            /* 指定主效应和交互效应项 */
```

运行程序 8.8 产生图 8.8 所示的结果。

结果分析:三种减肥药物对两组受试者的减肥平均效果没有差异。

H0:三种减肥药物对两组受试者的减肥平均效果(平均体重增加)没有显著差异。

根据因素的不同 GLM 过程输出两种形式的离差平方和:

TYPEⅠ SS:是按累积效应(有交互项)输出的离差平方和。如果有绝对的把握将所有的因素按主次顺序(先为主效应,后为交互效应)出现在 MODEL 语句中,则选择

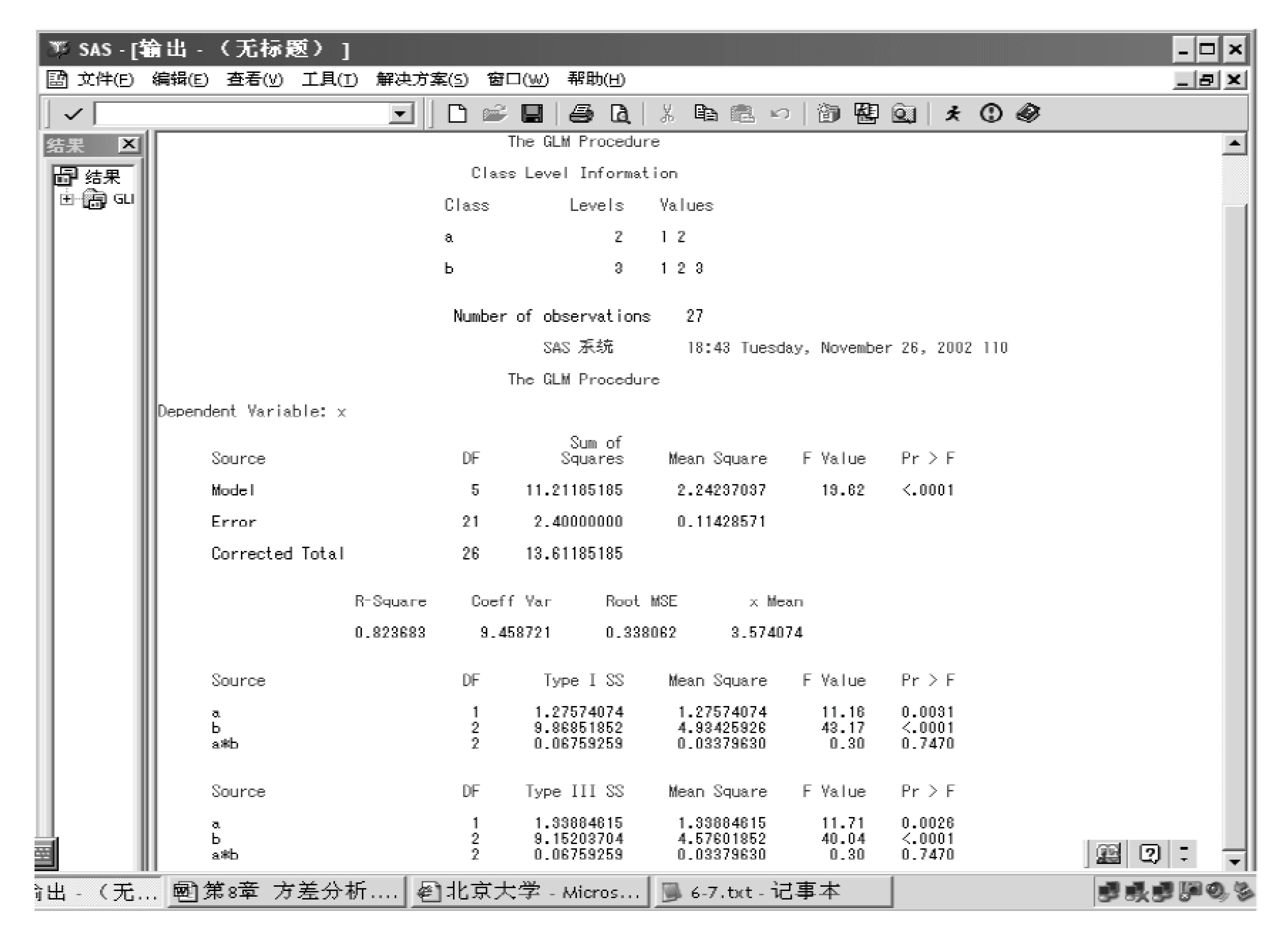


图 8.8　三种减肥药物对两组受试者的减肥效果的差异性比较

TYPE Ⅰ SS。否则应选择 TYPE Ⅲ SS。

TYPE Ⅲ SS：只有主效应时，按独立效应输出。

本例选择 TYPE Ⅰ SS。

如图 8.8 所示：模型的 F 值为 19.62，其概率小于 0.0001，必然小于 α 值 0.05，所以有足够的理由拒绝 H0。表明三种减肥药物对两组受试者的减肥平均效果(平均体重增加)有显著差异。

既然有差异，可进一步观察各组的情形。

从图 8.8 看，因素 a 的 F 值为 11.16、其概率为 0.0031<α 值 0.05，所以有足够的理由拒绝 H0，表明因素 a(男女)两组的平均体重有显著差异。

因素 b 的 F 值为 43.17、其概率为 0.0001<α 值 0.05，所以有足够的理由拒绝 H0，表明因素 b(三组药物)对平均体重有显著不同的疗效。

但不存在 a * b 的交互作用。

8.3　协方差分析

协方差分析是线性回归与方差分析的结合使用。它先通过回归关系删除自变量不等的影响，然后进行方差分析。例如，在减肥试验中，先用线性回归找出初始体重与新增体重的关系(或广告量与销售量的关系)，再计算当初始体重调整为相等时的各组平均新增体重，然后用协方差分析检验“调整的各组平均新增体重”有无差异。

协方差分析适合于本章开头所述的 4 种数据，即：

- 完全随机设计数据的协方差分析(即单因素协方差分析)
- 随机区组数据的协方差分析(即双因素协方差分析)
- 拉丁方设计数据的协方差分析(即三因素协方差分析)
- R * C 析因设计数据协方差分析(有交互作用的协方差分析)

8.3.1 GLM过程命令

1. GLM 的命令格式

```
PROC GLM;
  CLASS v1 v2;                    /* v1 和 v2 为分类变量 */
  MODEL Y= v1 v2/SOLUTION;        /*  Y 为数字型因变量,SOLUTION 要求计算参数估计值 */
  LSMEANS v4/选项;                /* 计算 v4 的最小二乘方均值 */
```

2. 格式说明

```
CLASS v1 v2;                      /* 定义 v1 和 v2 为分类变量 */
MODEL Y= v1 v2/SOLUTION;          /* 指定因变量 Y 和协变量 v1 与 v2。协变量即是自变量 */
LSMEANS v4/选项;
```

选项如下:

E:最小均方的估计函数
E=效应:用模型中的某一项作为标准误差项
STDERR:最小均方的标准误差
PDIFF:输出 H0 的检验值,包括置信区间
SINGULAR=n:对估计性检验的调整

8.3.2 用 GLM做协方差分析

已知健康人的甲胎蛋白含量为 0~20ng/ml。现用两种药物治疗(8+5)例受试者,两个月后测得甲胎蛋白含量如表 8.9 所示。试检验两种药物对甲胎蛋白含量的平均效用。

表 8.9 检验两种药物对甲胎蛋白含量的平均效用

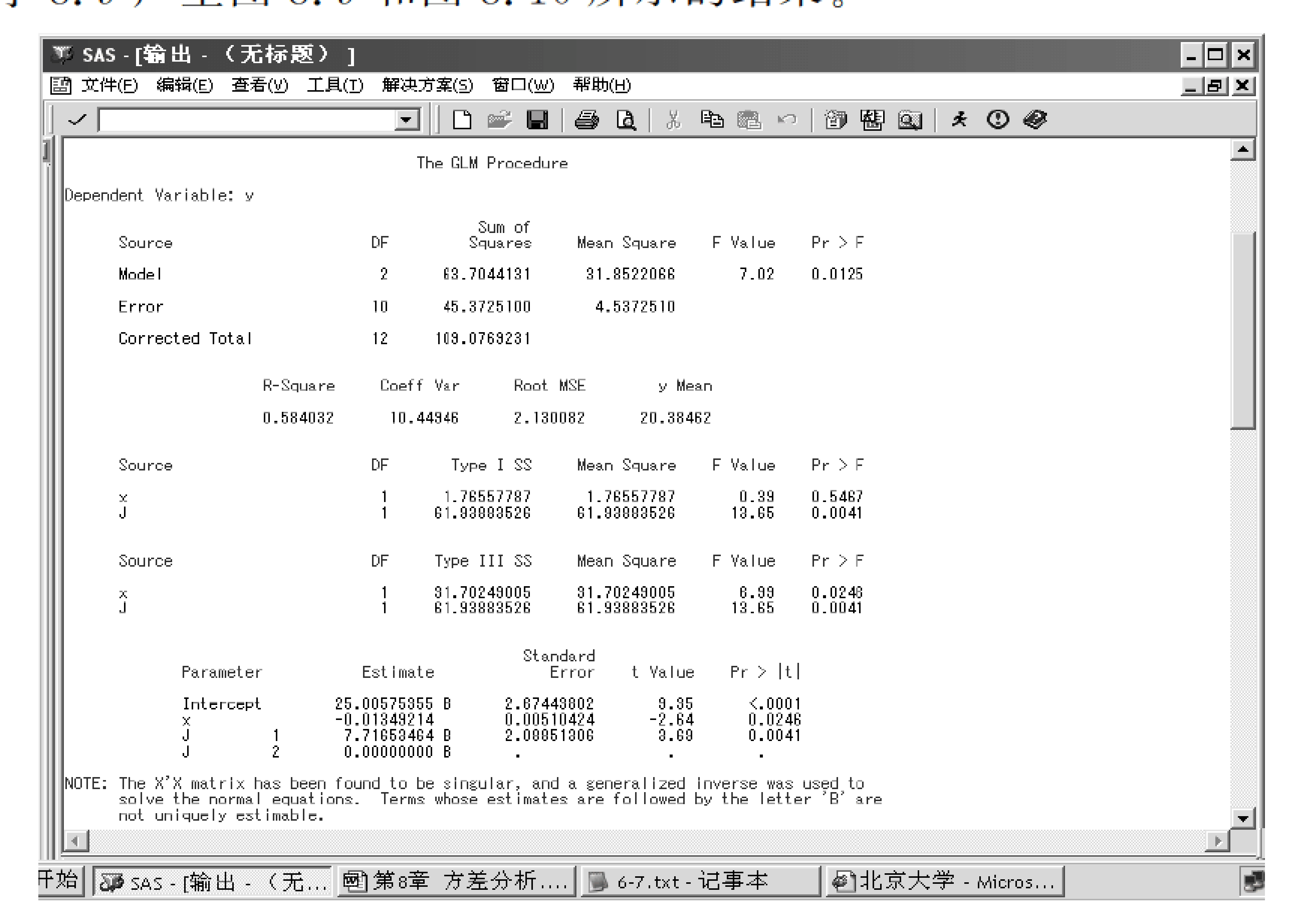
甲药		乙药	
药量(克)	甲胎蛋白量	药量(克)	甲胎蛋白量
580	25	530	22
680	23	500	20
700	25	480	18
800	22	450	15
900	21	488	17
1000	20		
980	18		
940	19		

解法：见程序 8.9。

程序 8.9：

```
DATA jj;
  DO J=1 TO 2;              /* j为甲乙两种药物 */
    INPUT n;
      DO I=1 TO n;          /* n为输入数据的次数 */
        INPUT x y@@;        /* x为药量,y为甲胎蛋白量 */
        OUTPUT;
      END;
    END;
  DROP I n;
CARDS;
8
580  25  680  23  700  25  800  22  900  21  1000  20  980  18  940  19
5
530  22  500  20  480  18  450  15  488  17
;
PROC GLM;
    CLASS j;                /* 定义 j为分类变量 */
    MODEL Y=x j/SOLUTION;   /* 指定 x为数字型因变量 */
LSMEANS j/STDERR;
    OUTPUT P=yp;
PROC PLOT;
    PLOT yp*x='*';
RUN;
```

运行程序 8.9 产生图 8.9 和图 8.10 所示的结果。



```
                              The GLM Procedure

Dependent Variable: y

                                        Sum of
     Source                      DF    Squares     Mean Square    F Value    Pr > F

     Model                        2   63.7044131    31.8522066       7.02    0.0125

     Error                       10   45.3725100     4.5372510

     Corrected Total             12  109.0769231

                    R-Square     Coeff Var      Root MSE        y Mean

                    0.584032      10.44946      2.130082      20.38462

     Source                      DF      Type I SS     Mean Square    F Value    Pr > F

     x                            1     1.76557787      1.76557787       0.39    0.5467
     J                            1    61.93883526     61.93883526      13.65    0.0041

     Source                      DF    Type III SS     Mean Square    F Value    Pr > F

     x                            1    31.70249005     31.70249005       6.99    0.0248
     J                            1    61.93883526     61.93883526      13.65    0.0041

                                                 Standard
          Parameter            Estimate             Error    t Value    Pr > |t|

          Intercept        25.00575355 B       2.67443802       9.35      <.0001
          x                -0.01349214         0.00510424      -2.64      0.0246
          J          1      7.71653464 B       2.08851306       3.69      0.0041
          J          2      0.00000000 B        .                .           .

NOTE: The X'X matrix has been found to be singular, and a generalized inverse was used to
      solve the normal equations.  Terms whose estimates are followed by the letter 'B' are
      not uniquely estimable.
```

图 8.9　检验两种药物对甲胎蛋白含量的平均效用

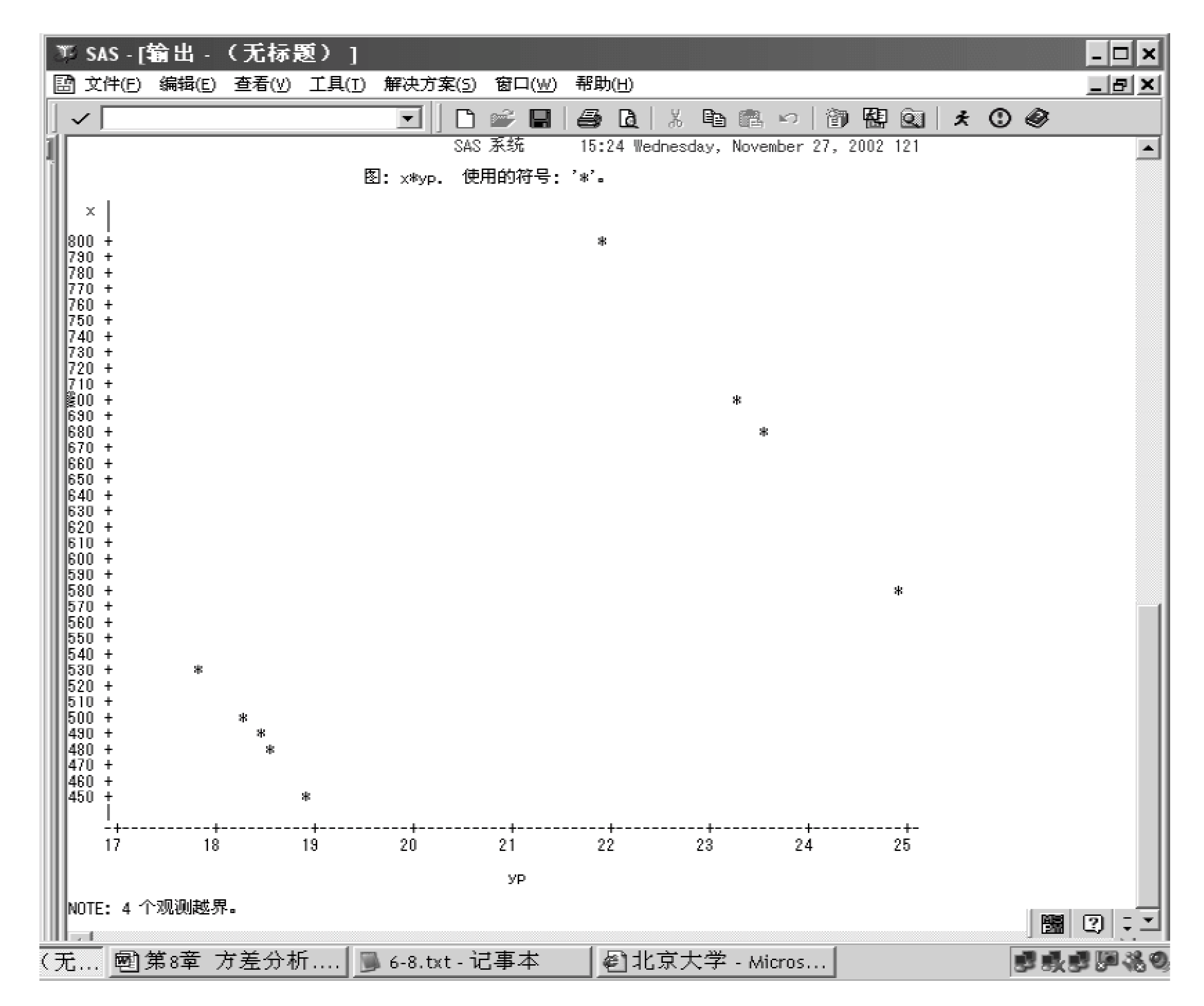


图 8.10 由 PLOT yp * x= ' * '产生的散点图

结果分析：

H0：检验两种药物对甲胎蛋白含量的平均效用没有差异。

检验：由图 8.9 看，模型的 F 值为 7.02，F 的概率值为 0.0125<α 值 0.05。所以有足够的理由拒绝 H0，表明药物对甲胎蛋白含量的平均效用有显著差异。可继续分析因素的作用。

由于因素 j 的 F 值为 13.65，其概率为 0.0041<α 值 0.05。所以两种药在药效调整后甲胎蛋白含量的平均效用有显著差异。

本例只有主效应，应该观察 TYPE Ⅲ SS 的离差平方和。对于因素 x(药量)，其 F 的概率为 0.0246<α 值 0.05。表明药物与甲胎蛋白含量之间关系显著，即 x 与 y 之间呈现线性关系。

所以调整后的直线模型为：甲胎蛋白含量=25.00575355-0.01349214 * 药量。

两组调整的均值分别如下：

Y(1)=7.71653464
Y(2)=0.00000000

因为 y 的检验概率 P 为 0.0041<α 值 0.05，所以 Y(1)一组调整后的平均甲胎蛋白含量显著差异。但是 Y(2)例外。

从图 8.10 看，药物与甲胎蛋白含量之间的线性关系显著。

为了练习多种编程方法，对程序 8.9 改用 INPUT 语句输入(见程序 8.10)。

程序 8.10：对程序 8.9 改用 INPUT 语句输入。

```
DATA jj2;
INPUT j x y @@;                    /* j为药组,x为药量,y为甲胎蛋白量 */
CARDS;
1 580 25 1 680 23 1 700 25 1 800 22 1 900 21 1 1000 20
1 980 18 1 940 19
2 530 22 2 500 20 2 480 18 2 450 15 2 488 17
;
PROC GLM;
    CLASS j;                       /* 定义 j为分类变量 */
    MODEL Y=x j/SOLUTION;          /* 指定 x为数字型因变量 */
LSMEANS j/STDERR;
    OUTPUT P=yp;
PROC PLOT;
    PLOT yp*x='*';                 /* 图点用"*"组成 */
RUN;
```

运行程序 8.10 同样可产生图 8.9 和图 8.10 所示的结果。

习　题　8

1. 常用的方差分析法有哪 4 种?

2. 均衡数据和非均衡数据各用什么过程命令进行分析?

3. 对血小板偏低者用 4 种不同的药物治疗后血小板的数据见程序 A.36。试用 PROC ANOVA 命令检验 4 种药物平均疗效有无显著差异。

程序 A.36:

```
DATA f2;
    DO i=1 TO 8;
      DO g=1 TO 4;
       INPUT Y @@;
       OUTPUT;
      END;
    END;
CARDS;
10.1  9.0   10.6  10.1
9.5   11.2  9.2   11.5
10.0  11.6  10.0  10.2
9.2   10.8  8.9   10.5
8.5   10.0  8.8   11.5
9.0   10.2  8.4   10.6
9.5   11.8  10.1  10.5
10.1  11.8  10.5  11.9
;
```

```
PROC ANOVA;
  CLASS i g;
  MODEL Y= i g;
  MEAN g/DUNNETT;      /* 增加两两比较的功能 */
```

4. GLM 过程的统计功能有哪些？

5. 试用 GLM 过程检验程序 A.37 中各组平均红细胞数目有无显著差异。

程序 A.37：

```
DATA RBC;
  DO J=1 TO 3;
  INPUT n;
    DO I=1 TO n;
     INPUT x @@;
     OUTPUT;
    END;
   END;
CARDS;
9
6.35 6.50 6.70 5.80 6.35 6.80 6.40 5.75 6.15
8
5.65 5.50 6.30 5.60 5.15 5.75 6.10 6.85
7
3.80 5.20 5.10 4.50 4.95 5.15 5.85
;
PROC GLM;
    CLASS j;            /* 定义 j 为分类变量 */
    MODEL x=j;          /* 指定 x 为数字型因变量 */
MEANS t1/SNL;           /* 用 SNK 的 Q 检验法检验 */
```

6. 已知健康人的甲胎蛋白含量为 0～20ng/ml。现用两种药物治疗(8+6)例受试者，两个月后测得甲胎蛋白含量如程序 A.38 所示。试检验两种药物对甲胎蛋白含量的平均效用。

程序 A.38：

```
DATA jj;
  DO J=1 TO 2;            /* j 为甲乙两种药物 */
   INPUT n;
    DO I=1 TO n;          /* n 为输入数据的次数 */
     INPUT x y@@;         /* x 为药量，y 为甲胎蛋白量 */
     OUTPUT;
    END;
   END;
  DROP I n;
CARDS;
```

```
8
580   25   680   23   700   25   800   22   900   21   1000   20   980   18   940   19
6
530   22   500   20   480   18   450   15   488   17   500   19
;
PROC GLM;
     CLASS j;                          /* 定义 j 为分类变量 */
     MODEL Y=x j/SOLUTION;             /* 指定 x 为数字型因变量 */
LSMEANS j/STDERR;
     OUTPUT P=yp;
PROC PLOT;
     PLOT yp*x='*';
RUN;
```

第 9 章 chapter 9

相关分析

相关分析的数据，基本上可以分为四对(类型)：①定类—定类(即：标称—标称)、②“定序—定序”(即：次序—次序)、③“定类或定序—定距”(即：非连续数据—区间)、④“定比—定比”(即：比例—比例)。

相关系数是一个“无单位”的系数，其值的大小表示线性相关程度的强弱。正负相关系数表示相关的方向为正或为负。通常用字母 r 表示相关系数，r 值的范围为：$-1\leqslant r\leqslant 1$。当 $r=0$ 时，说明两个变量不一定无关，而是呈现着不规则的变化。

SAS 中的线性相关主要包括：皮尔逊(Pearson)的积差相关、斯皮尔曼(Spearman)的等级相关、肯氏(Kendall)的等级相关以及 Hoeffding 的 D 相关系数。

9.1 数据的 4 种测量水平

在讨论变量的相关之前，只有弄清变量的层次才能准确地应用相关公式求出变量间的相关程度。变量的层次有如下 4 种。

(1) 标称变量(Nominal Variable)：国内称之为“定类变量”或“名义变量”。其变量值无等级、次序之分，而仅仅是表示东西南北、张三李四或男性女性之别，所以被国际上通称为标称变量。这种变量是 4 种变量层次中最低的一种。

(2) 次序变量(Ordinal Variable)：国内称之为“定序”变量。其值有等级高低、次序先后之分。例如分组后的“教育水平”变量，其值可以是小学、中学、大学程度等层次。又例如，职务这个变量，其值可以是科员、科长、处长等级别。这种变量是 4 种变量层次中次低的一种。

(3) 区间变量(Internal Variable)：国内称之为“定距”变量。其变量值既有等级和次序之分，又可测量出某区间的值，例如温度或气温，不但具有次序变量的性质，而且还可以反映今天比昨天气温高出多少度。这种区间变量的层次高于次序变量，居于第二位。区间变量只有温度(气温)、海拔高度等极个别的变量。

(4) 比例变量(Ratio Variable)：国内称之为“定比”变量，其变量值既具有次序变量和区间变量的性质，还存在一个有意义的“零点”。例如，甲高 2 米，乙高 1 米，甲则是乙的两倍高度。又例如一个人的血压不能是 0mm 水银柱，这个有意义的 0 在这个变量中起着质的变化。这样的变量是比例变量(定比变量)，它是 4 种变量层次中的最高级。

以下的章节将对这4种变量的水平(level)及其测量方法,各举出一个例子加以比较。

9.2 皮尔逊积差相关

当两个分析的变量,其值为比例—比例变量的层次时,不仅可以研究其相关程度,而且还可以研究其相关方向是线性或非线性的。本节主要是采用皮尔逊相关公式来研究变量之间是否存在线性相关。这种相关应考虑到 x 与 y 变量值本身的大小。

9.2.1 皮尔逊相关系数CORR的计算公式

皮尔逊相关系数CORR的计算公式如下:

$$R = \sum xy/(nS_xS_y) = \sum (X-\bar{X})(Y-\bar{Y})/(nS_xS_y) \tag{9.1}$$

公式(9.1)中,

S_x是变量 X 的标准偏差。

S_y是变量 Y 的标准偏差。

$\bar{X}$ 是变量 X 的平均值。

$\bar{Y}$ 是变量 Y 的平均值。

9.2.2 皮尔逊相关系数的测量

下面图9.1中的变量及其数据,是1991年北京大学郭崇德教授关于北京市等大中城市社区服务研究中《居民调查》的数据,笔者从几千名受访者中随机抽取66名,以计算人均月结余Vio(定比数据水平)与人均居住面积Av8f(定比变量水平)之间的皮尔逊(Pearson)相关系数。

程序9.1:

```
DATA sq;
INPUT Id 1-2 caseid 3-5 age 6-7 sex 8
      edc 9 wk 10 fm 11 V6f 12 V7f 13-14
      v8f 15-16 V9f 17 v10a 18 v10b 19 v10c 20 v10d 21 vi 22-24 vo 25-27;
IF age=0|sex=0|fm=0|edc=0|wk=0|V6F=0|v7f=0|v8f=0|V9F=0 THEN DELETE;
IF vi=0|vo=0|v7f=0|v8f=0 then delete;
   oi=vo/vi; vio=vi-vo; av8f=v8f/v7f;av=vi/v7f;
LABEL     sex='户主性别' edc='文化程度' wk='具体工作'
        ID='区与街道代号' caseid='问卷号'
        fm='婚姻状况' v6f='几代人?' V7f='总人数' V8f='居住面积' V9f='住房类型'
        v10a='煤气'  v10b='卫生间'  v10c='暖气'  v10d='自来水'
        Vi='月收入:元' Vo='月开支:元';
CARDS;
1100141152230528100017133500110803010101010102003443222120001000001000
```

```
1100269250230542301113002500100004020201000102031111113020001000102000
1100374156130644110016504501000003020100000102081104433110004000000000
1103067142230636200013202000100002030201010702051132333130003000001000
1103156142230625210013603001100802000300000102001133333130002000021000
1103234243230736110013903900100000000000000102003202213120002000021060
1103378136230930210017004000101204010101000102002344332130000000001050
1103455136230733210016205501101103000000020102001103333120001000011000
1103558136220324100012702400100000000000000102003242243120010100100050
1106725233210215210114003000000000000000000000333443312000000001000
1106840242220311210013003000000803000002000000003304342000001000010000
1106964120222021821001200170000000300000000000000111111102000001000000
1107053155220330110013002000001102000000000102092253432130002000031000
1107162216230742210113502001000005040000000100001153333320000000041060
1201358246200040300014003001101000000000000102003402111121110111101010
1201459152220433411116502800100801020100020102052342221111110000000070
1201561226230454410111200501100010030000000105002442222110000001310060
12016302422306354101140040011008020102011001020835333331120003000101020
1201761226230750311116504001100006030000000103003343433120100100001050
1201823146220440311116003000100901010100000102032254433120000000000010
1201967216430437301115504000100004020000000000002222222120000000121000
1202067154230424311114000000100003020100000102063452322120000000010000
1202129155230800311113503001101201000203000102051231122110000000010020
1202263230220338411113003500100004030300000102082222222122100101001000
1204655216240425300113833001100006020000040107091121122120000000151010
1204777120210236311112832000100000020100050102093231111120050304100040
1204837232220342311113500000000102000000000108001332223120001000111000
12049552262102264111125017000000050300010001070022322233110000100111060
1207330141220334311113001501000000000000000102003332222020001000011000
1207459232230547411116004001100007020100000105073204433120002000020000
1207552242020355411114002000000002020200000102091101111100000000000000
1207628132220325311113002001000801020200000104002242222120001000010020
1207726252120325311115004000000801000300000102001121122120001000111010
1300365250231060201019999991100008030100000102091141133120000000000000
1300465245220450411115302000100204020202000102081133333120000000001000
13019352402311602100150040011008000000002000000111141221120003000211000
1302063231231221210018007000100003000000000102003000000300001000010000
1302163132230430110013002001000003000100000206092351131021000300101040
1302223156120618200006006001000000000300000207001111111120001000000050
1302354220231620210001441001100010000100000102003353333120003000001000
1302465254230438411114002500100003000200020102052243333120000000000000
1302561131230343411112801000000020304000000000023533331000000000000000
1302660244230660411112001801100005000002000102003344343120020011201030
1302763120210222210013702000100601000000000100002353333120000000011000
1302871152200040210012001500000002010001000102072332222110000000001020
```

```
130296825443055121001240200010020603010000010708234333313000400000000
130513224123052321000480300110000501010102010208111111112000000110000
130528012623052721001500400010000402010017010207334333233000000001000
130533815222041521001100080010080401010101010000121111102000100001103 0
130543124123061211000600350110000000000000010211125433210000400005100 0
130556322643073020000700500110001001000000010200345333210000000000100 0
130564214122042820001500360010110202010107010811234111112000000001100 0
130776922244043231111350250100060401010100010204235333312070000000102 0
130785225212033741111350250000080200030102010211233333212000000000100 0
130796020023040011000500350100000901010300000000235323312000000000100 0
130804925422031910001400250000080302030008010200135333333000100000100 0
130812315612043621001110440000008000000000010204345554433000200002100 0
130825711123062420001000000011080200000000050000120222302000000000102 0
140057122643041631001450300010000402010201010304111111111000041010101 0
140307121643041820000240240010000602000004011100333322312100010010104 0
140315013622041721001235235010000000000000000000345332433000200010101 0
140326022623041611000142130010000404000002010700445513312000400005100 0
140713813322032011000400350000080300000203010400345352212000000010104 0
140725413623072020001400280111120102010000010200124334212023010031100 0
140735721133062031001800400100000800000000010300233532312000000011105 0
140744515122031020001350300000080300010101010200335553312000000000103 0
RUN;
PROC FORMAT;
VALUE sex 1='1' 2='0';
VALUE sex 1='男' 2='女';
VALUE edc 1='文盲' 2='小学' 3='初中'
          4='高中或中专' 5='大专以上';
VALUE id 11='东城区东四居委会' 12='东城区安德里居委会'
      13='东城区赵家楼居委会' 14='东城区南小街居委会'
      31='宣武区牛街街道办' 32='宣武区广内街道办' 33='宣武区天桥街道办'
      41='海淀区中关村七区' 42='北太平庄蓟门里居委会';
VALUE wk 1='工人' 2='干部 (职员)' 3='商业服务人员'
          4='教员' 5='科研或技术人员' 6='其他';
VALUE fm 1='未婚' 2='已婚' 3='离婚' 4='丧偶';
VALUE V9F 1='四合院' 2='普通平房' 3='一般楼房' 4='高层塔楼'
          5='其他活动房等';
VALUE av8f LOW-5='1' 5-6='2' 6-7='3' 7-8='4'
          8-HIGH='5';
VALUE av8f 1='人均 5平米以下' 2='人均 5-6平米以内'
          3='人均 6-7平米以内' 4='人均 7-8平米以内'
          5='人均 8平米以上';
VALUE vio LOW--200='0' -200--100='1' -100-0='2'
          0-00='3' 100-200='4' 200-HIGH='5';
VALUE Vio 1='月结余:-200至-100元' 2='月结余:-100至 00元'
          3='月结余:00至 100元' 4='月结余:100至 200元'
          5='月结余:200元以上';
PROC CORR; VAR vio av8f;
RUN;
```

进入 Analyst(分析家)对话框的步骤如下:

（1）在图 9.1 所示的程序编辑器中编辑程序 9.1，按 F8 键运行程序 9.1 及其数据，产生 SAS 数据集 Work. sq。

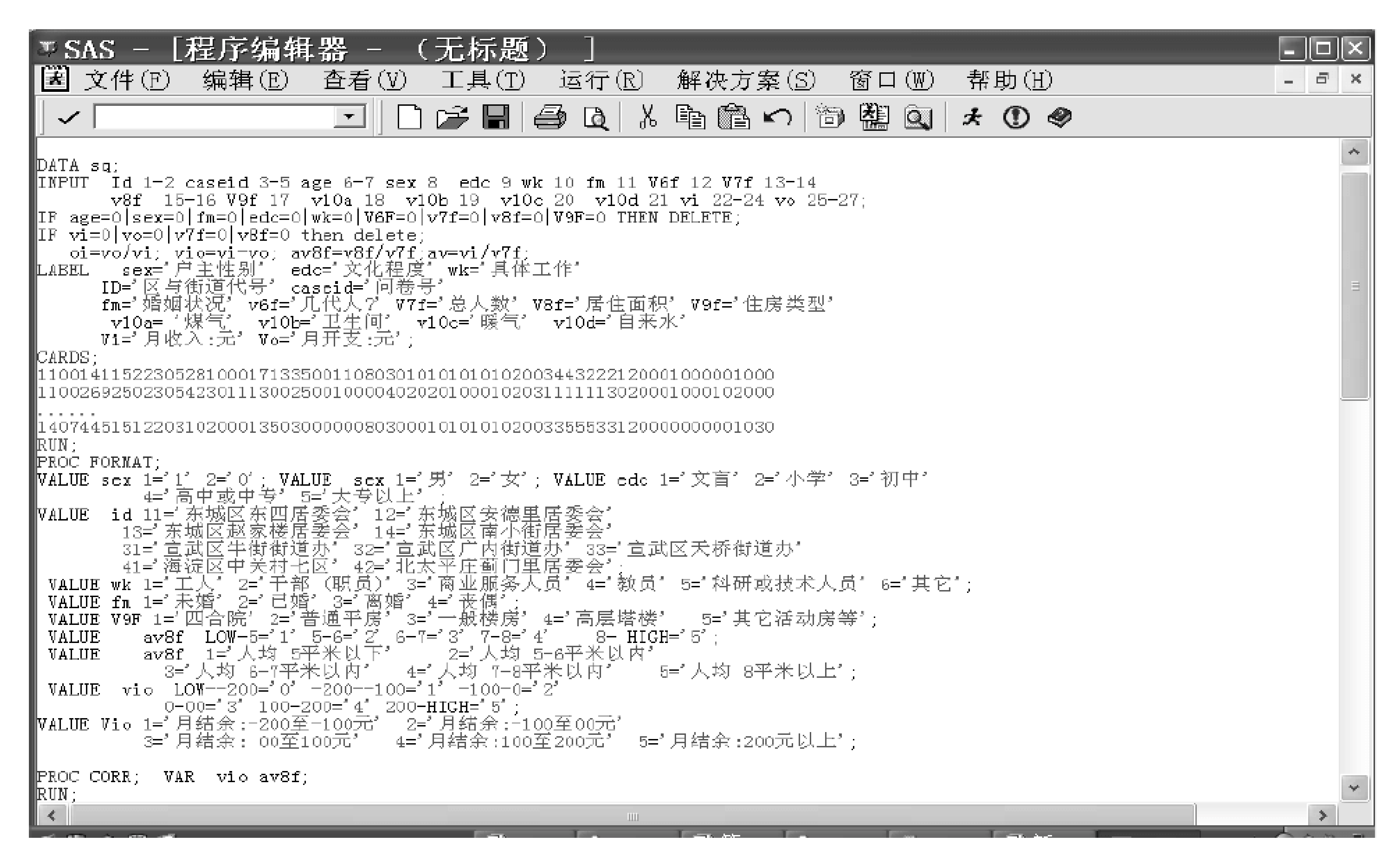


图 9.1 编辑程序和数据文件

（2）选择 SAS 主菜单中的“解决方案”（英文版为 Solutions）→“分析”（Analysis）命令，选择图 9.2 所示的“分析家”（Analyst）命令。



图 9.2 “分析家”（Analyst）的菜单位置

（3）选择“分析家”（Analyst）→“文件”（File）→“按 SAS 名称打开”（Open By SAS

Name)→Work 命令和按钮，进入图 9.3。



图 9.3　Work.sq2 工作文件

(4) 选择文件名 Sq2 后单击“确定”(或 OK)按钮，展示 Work.sq2 数据集的内容，见图 9.4。

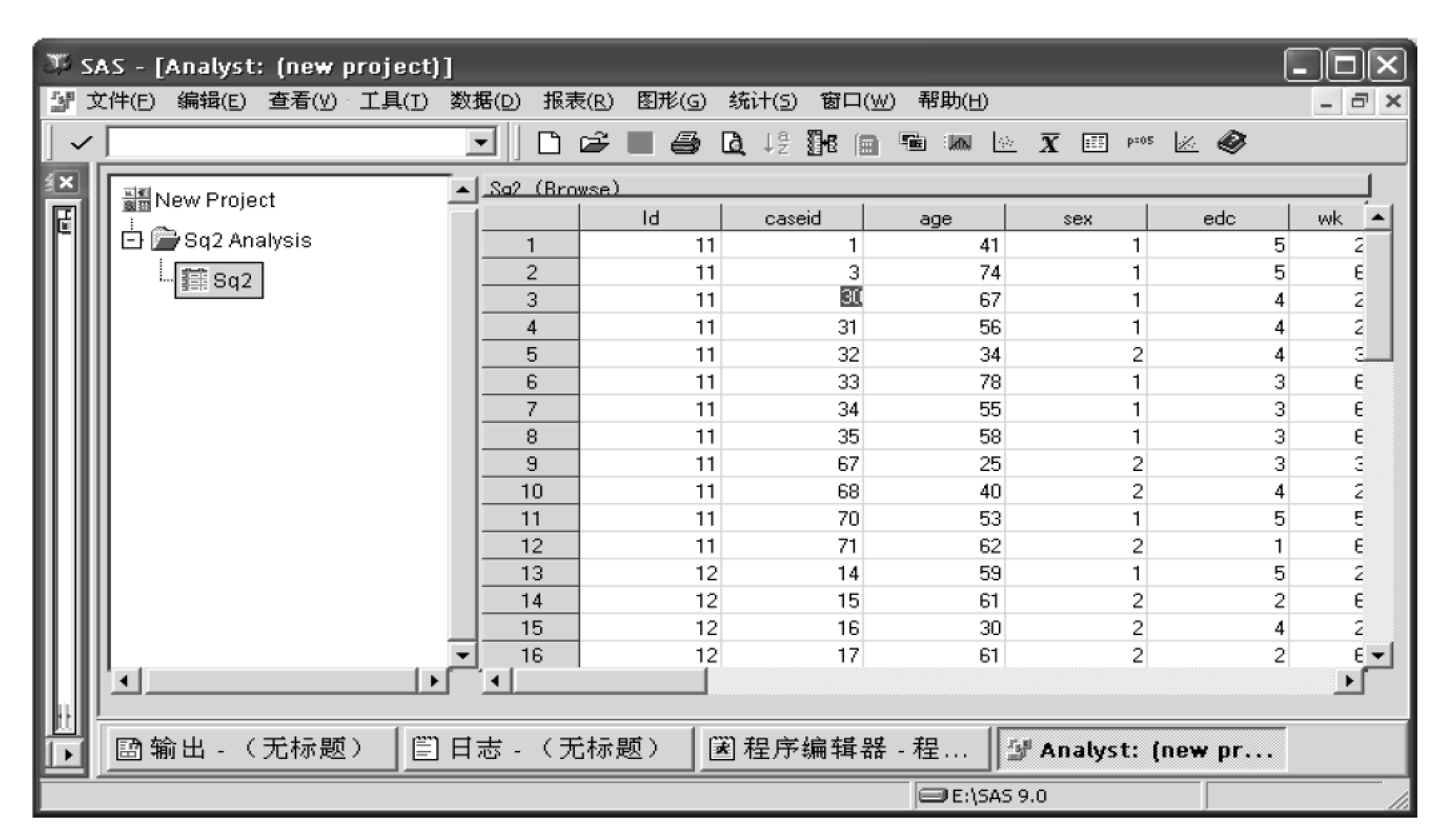
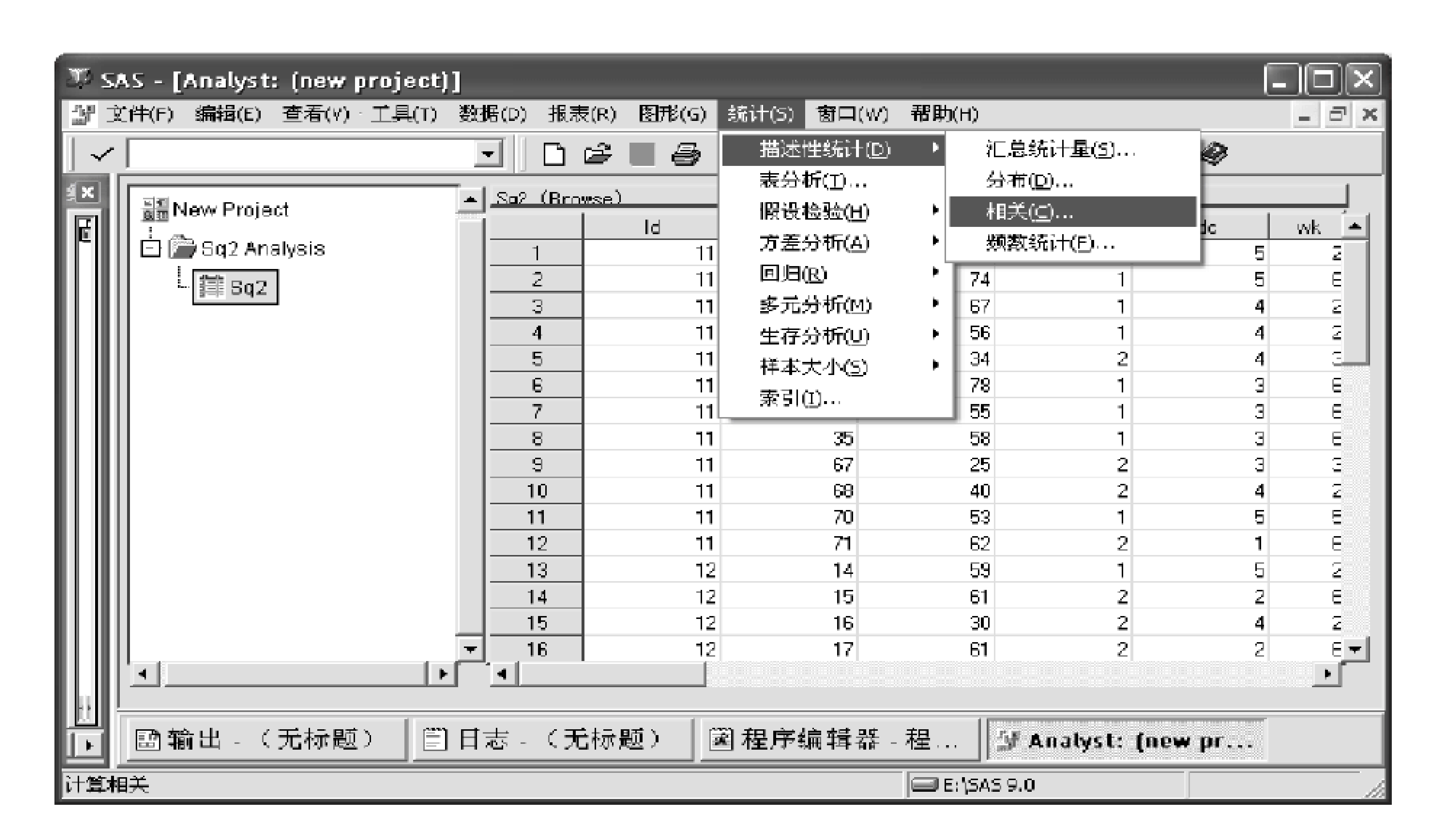
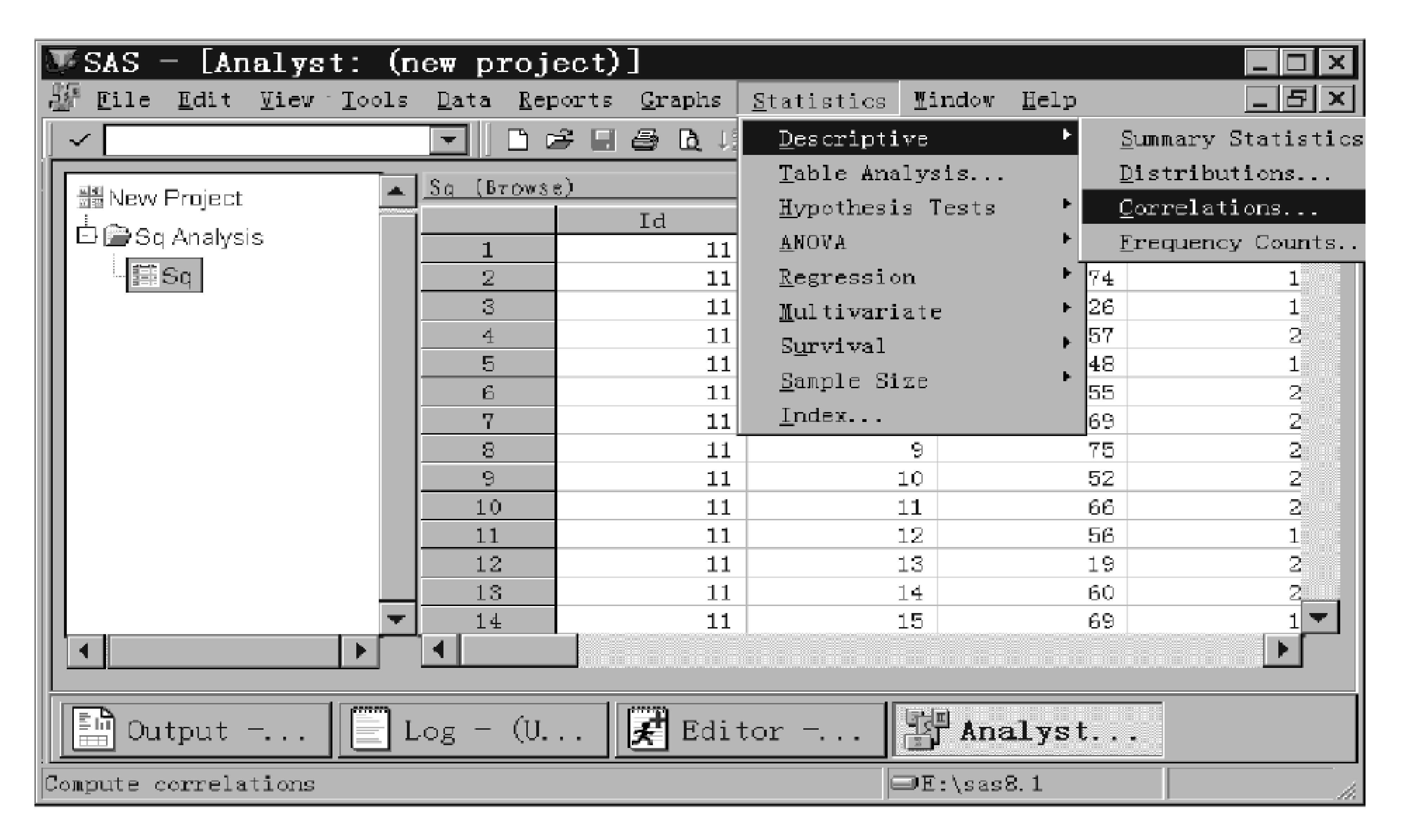


图 9.4　Work.sq2 文件的内容(部分)

(5) 选择“统计”(Statistics)→“描述性统计”(Descriptives)命令，展示如图 9.5 所示。



(a) SAS 9.0 以上版本的对话框



(b) SAS 8e 及以下版本的对话框

图 9.5 Correlations 的菜单位置

(6) 选择“相关”(Correlations)命令,进入图 9.6 并设置变量。

(7) 单击 Options,进入图 9.7 选择皮尔逊相关系数 Pearson。

(8) 单击两次 OK 按钮,输出图 9.8 所示的结果。

9.2.3 皮尔逊相关系数 CORR 的分析

人均月结余与人均居住面积的皮尔逊积差相关系数如图 9.8 所示。

当 H0: Rho=0 时,Prob > |r| under: 表示原假设相关系数 R 为 0 时的概率值 P。

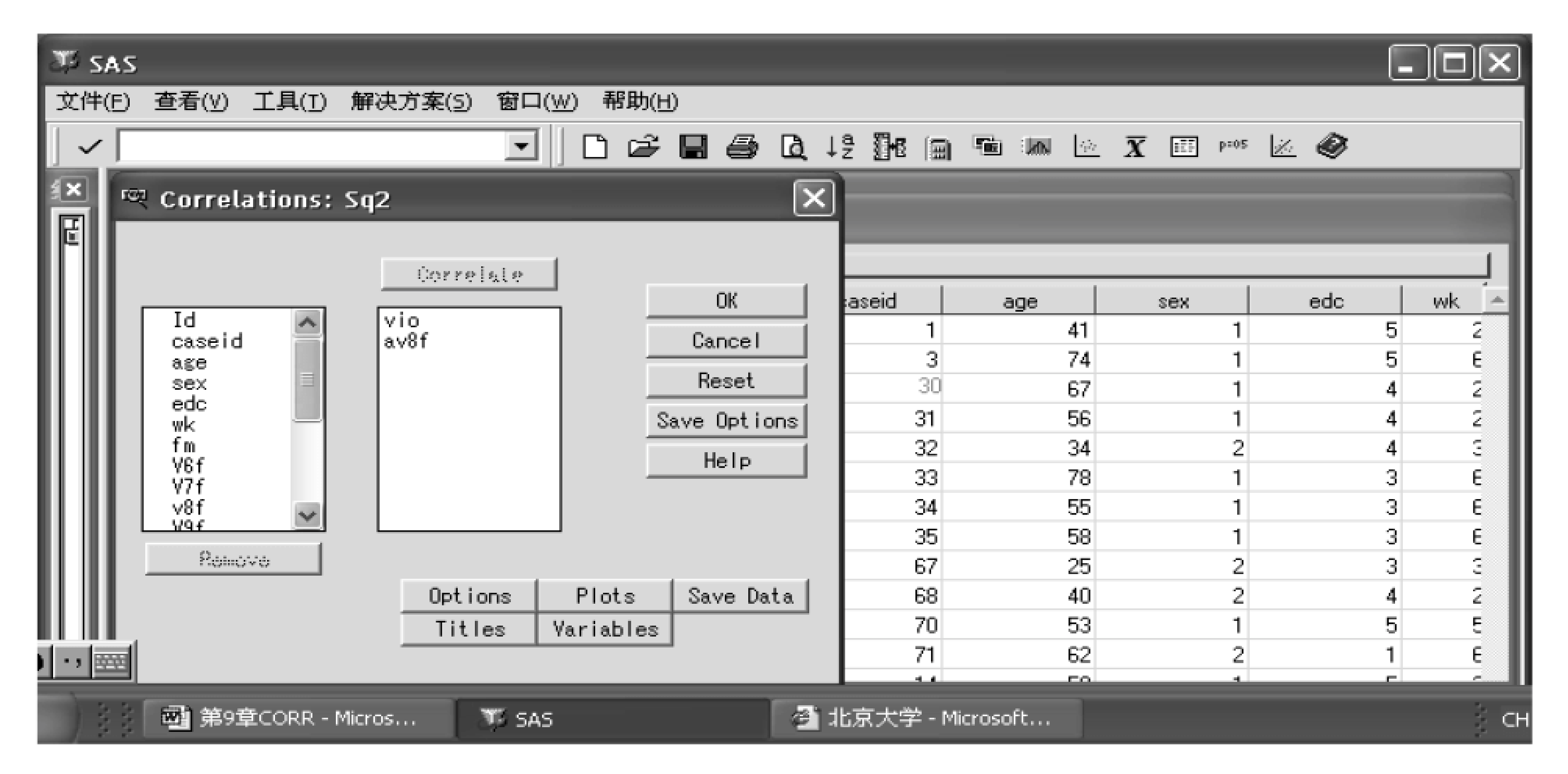


图 9.6　选择人均月结余与人均居住面积两个变量

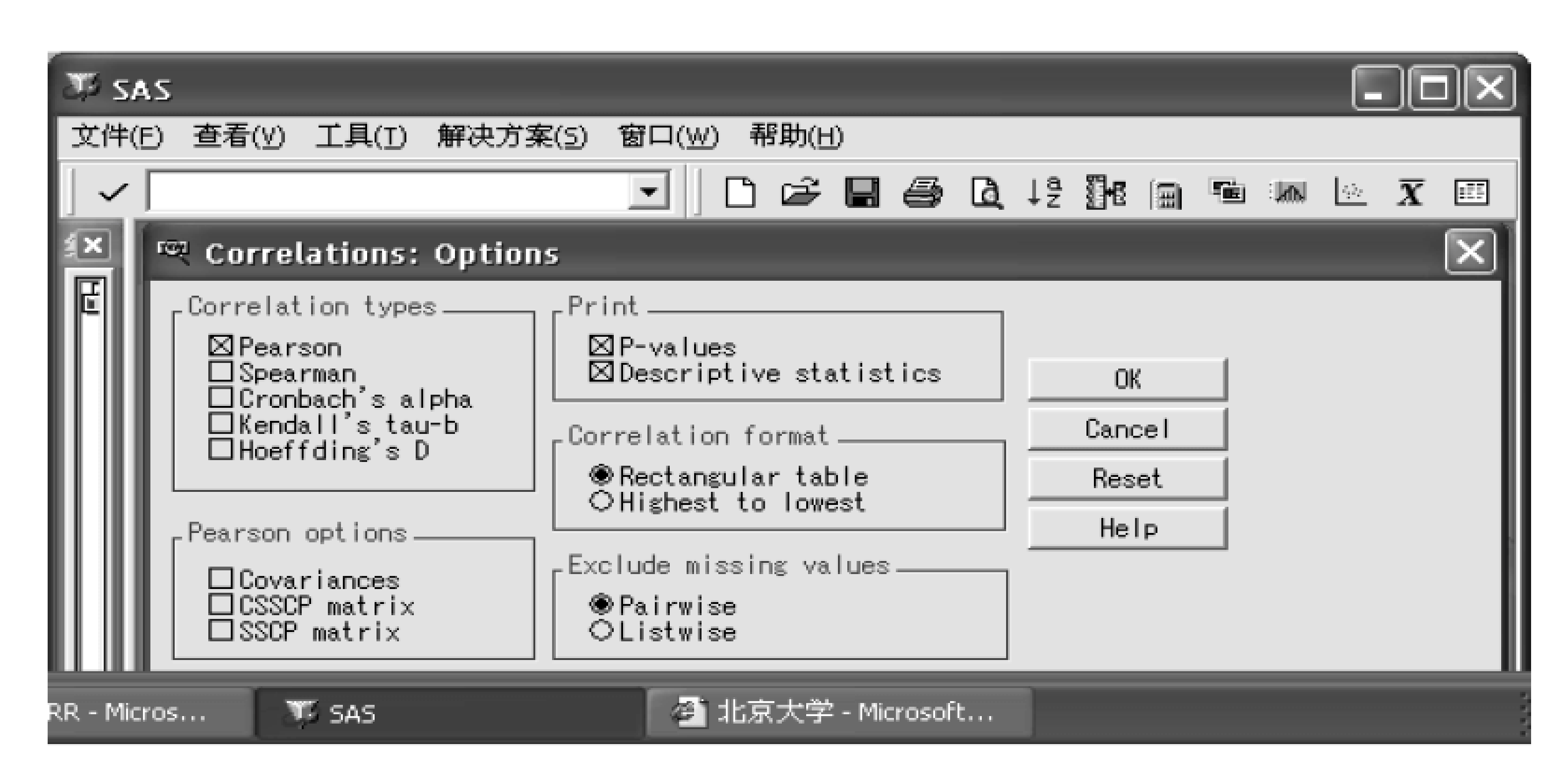


图 9.7　选择皮尔逊相关系数

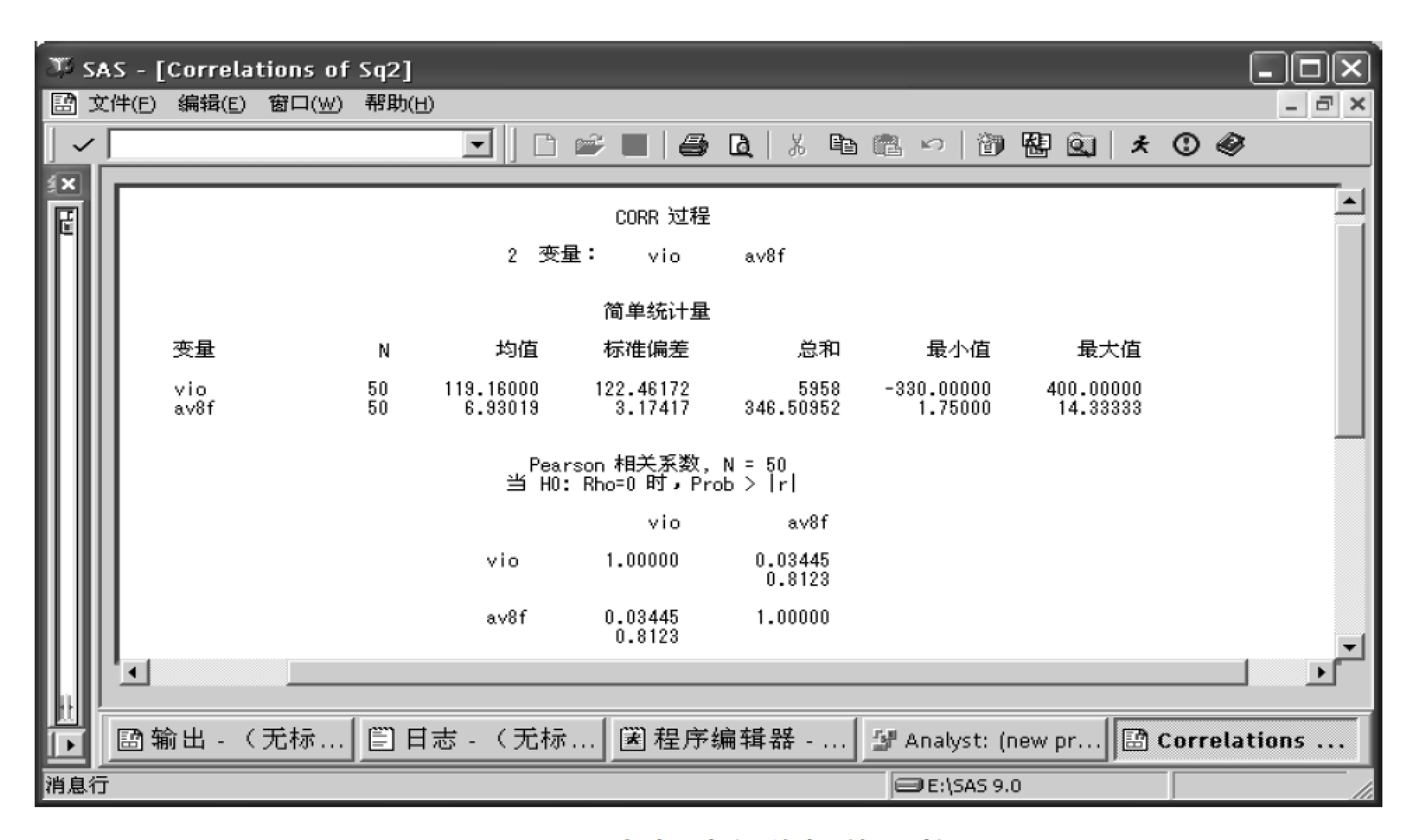


图 9.8　皮尔逊积差相关系数

由于 P 值=0.8123,大于 α 值 0.05,所以相关系数不显著。

又因为 r=0.03445,说明“人均月结余与人均居住面积”的皮尔逊积差相关系数极其小,几乎无关。建议将样本量增多到几百个,然后再试验。

9.3 皮尔逊二分“点—距”相关

当一个变量是 0 与 1 编码的二分变量(如性别编码为:1 男、0 女),另一个是定距以上变量时(如成绩、收入、奖金、血压等),如果没有 Eta 系数供选择时,也可试用上面介绍的皮尔逊积差相关系数测定是否相关。但慎用!

1. 统计步骤:参阅 9.2.2 节的步骤(1)~(6)。
2. 相关系数选择:参阅 9.2.2 节的步骤(7),单击 Pearson 相关统计量。
3. 结果分析:参阅 9.2.3 节。

9.4 肯氏(Kendall)等级相关 τ_b

肯氏的等级相关系数 τ_b,与斯皮尔曼的等级相关系数 R,具有相同的用法,即:用于次序—次序变量的相关测量。例如,如果计算 X、Y 两个变量的相关时,则只考虑 X 与 Y 变量值的大小顺序(等级),而不考虑变量值本身的大小。

肯氏等级相关与斯皮尔曼等级相关,二者在计算公式上有所区别。肯氏的计算公式是基于:对 X、Y 这一对变量值“和谐对”(Concordant)占多少,“不和谐对”(Discordant)占多少,二者之差作为分子、并以样本规模 n 所形成的总对数 $C_n^2=n(n-1)/2$ 作为分母而计算相关系数 τ_b 的。

和谐对,是指在同一个 OBS(观察值)中,X、Y 两个变量值都高于(或都低于)另一个 OBS(观察值)中的相应变量值。和谐对用 P 表示。

不和谐对,是指在某一个 OBS(观察值)中,变量 X 的值大于另一个 OBS(观察值)中的 X 值;而 Y 变量的值正好相反。它用字母 Q 表示。

肯氏公式还用到相持(Tied)概念。即:一对观察值 OBS(即个案)中,若有一个变量(或两个变量)的值对应相等,则称该对观察值 OBS(问卷)是相持的。

相持还分为相持在 X 变量上(记为 Tx),或相持在 Y 变量上(记为 Ty)。例如甲身高 1.7 米、体重 65 公斤,乙也身高 1.7 米、但体重 60 公斤,这时甲乙两个观察值(人)是相持在“身高”变量上。

因此,肯氏相关系数 τ_b 的计算公式为:

$$\tau_b=\frac{(P-Q)}{\sqrt{(P+Q+Tx)*(P+Q+Ty)}} \tag{9.2}$$

肯氏相关系数的计算公式还有其他种,因为不常用而省略。

9.4.1 计算肯氏等级相关系数的数据

下面图 9.9 中的变量及其数据是 1991 年某些大中城市社区服务研究中关于《居民



图 9.9　居民调查数据中的“等级—等级”数据

调查》的数据。

程序 9.2：见图 9.9。

其中：(1) Vio 是人均月结余，分组为 6 个等级：

LOW-200 元＝'0'　　－200 元-100 元＝'1'　　－100 元-0 元＝'2'

0 元-00 元＝'3'　　100 元-200 元＝'4'　　200 元-HIGH＝'5'；

(2) Av8f 是人均居住面积，分组为 6 个等级：

LOW～60 米^2＝'0'　　60 米^2～70 米^2＝'1'　　70 米^2～80 米^2＝'2'

80 米^2～90 米^2＝'3'　　90 米^2～100 米^2＝'4'　　100 米^2～ HIGH＝'5'；

要求：计算人均月结余 Vio(等级定序数据水平)与人均居住面积 Av8f(等级定序数据水平)之间的肯氏相关系数 τ_b。

解法见 9.4.2 节。

9.4.2　通过 Analyst 中的对话框计算肯氏相关系数 τ_b

操作步骤如下：

(1) 运行图 9.9 中的程序生成数据集 sq2。

(2) 选择 SAS 主菜单中的“解决方案”(英文版为 Solutions)→“分析”(Analysis)命令，鼠标指针移到图 9.10 带有阴影标记的“分析家”(Analyst)命令上。

(3) 选择“分析家”(Analyst)→“文件”(File)→“按 SAS 名称打开”(Open By SAS Name)→Work 命令和按钮，进入图 9.11。

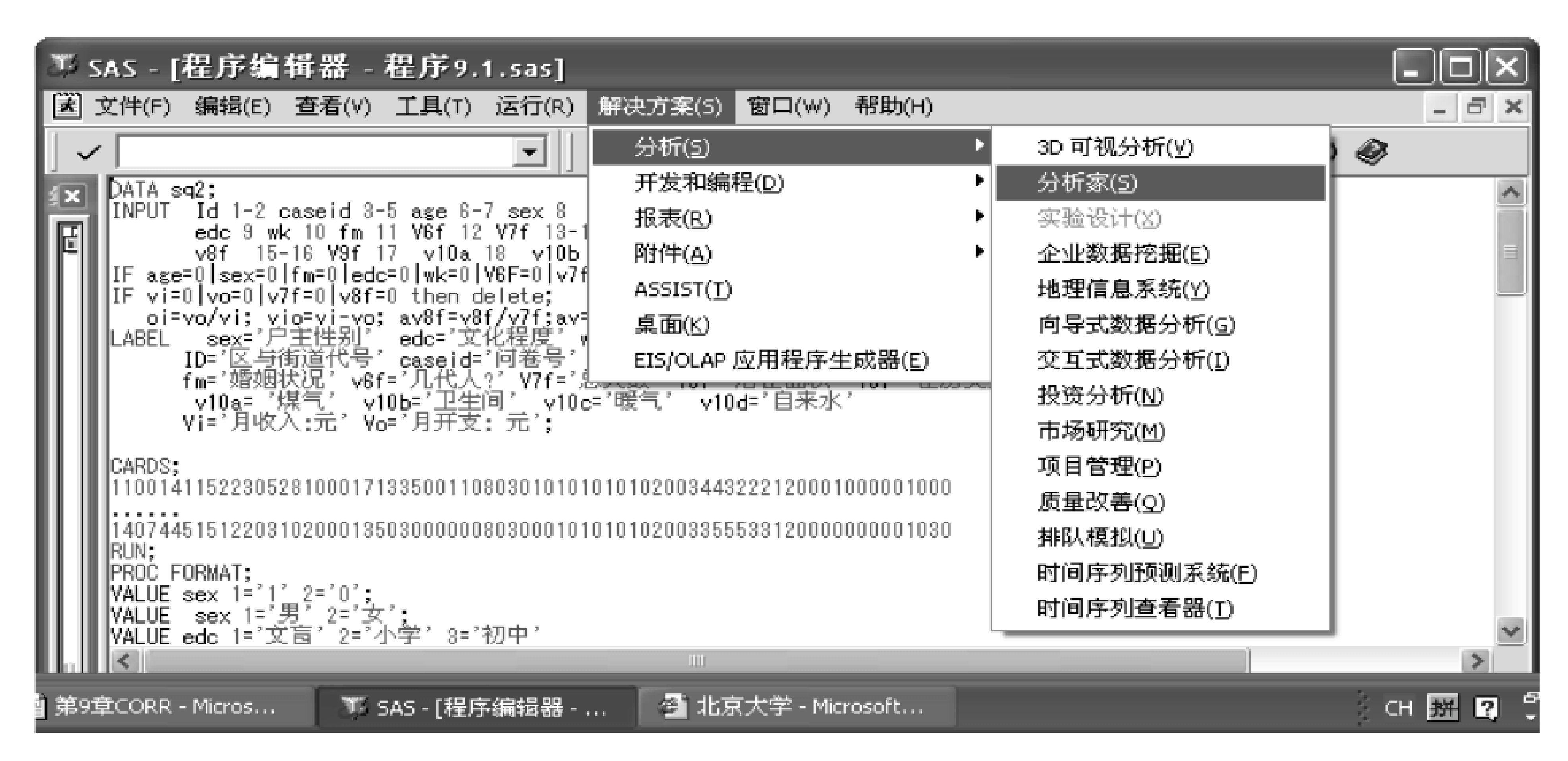


图 9.10　分析家(Analyst)的菜单位置



图 9.11　Work.sq2 工作文件

(4) 选择文件名 Sq2 后单击 OK 按钮,展示 Work.sq3 数据集的内容,见图 9.12。

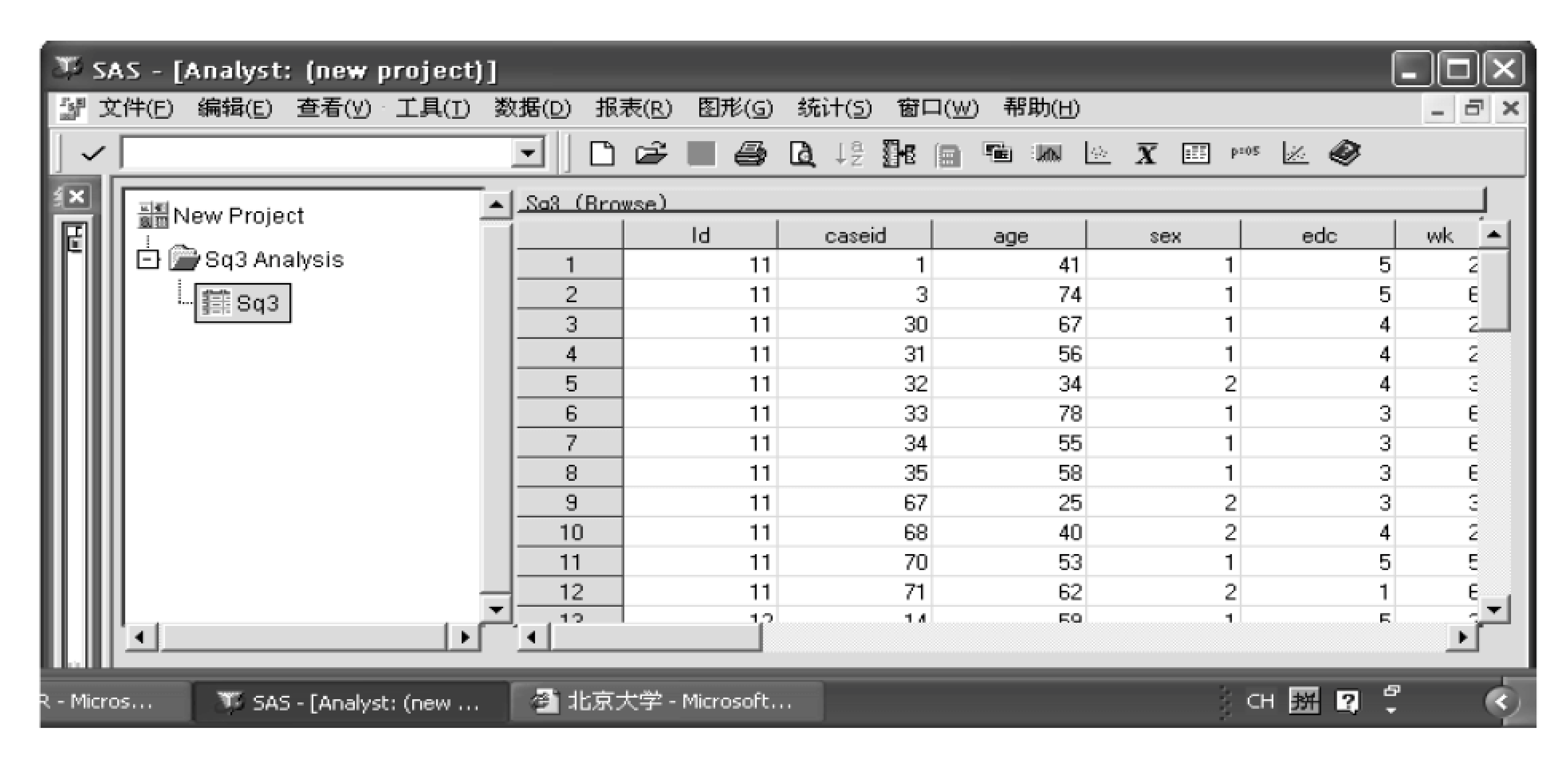


图 9.12　Work.sq2 文件的内容(部分)

(5) 选择"统计"(Statistics)→"描述性统计"(Descriptives)命令,展示图 9.13。



图 9.13 Correlations 的菜单位置

(6) 选择"相关"(Correlations)命令,进入图 9.14 并设置变量。

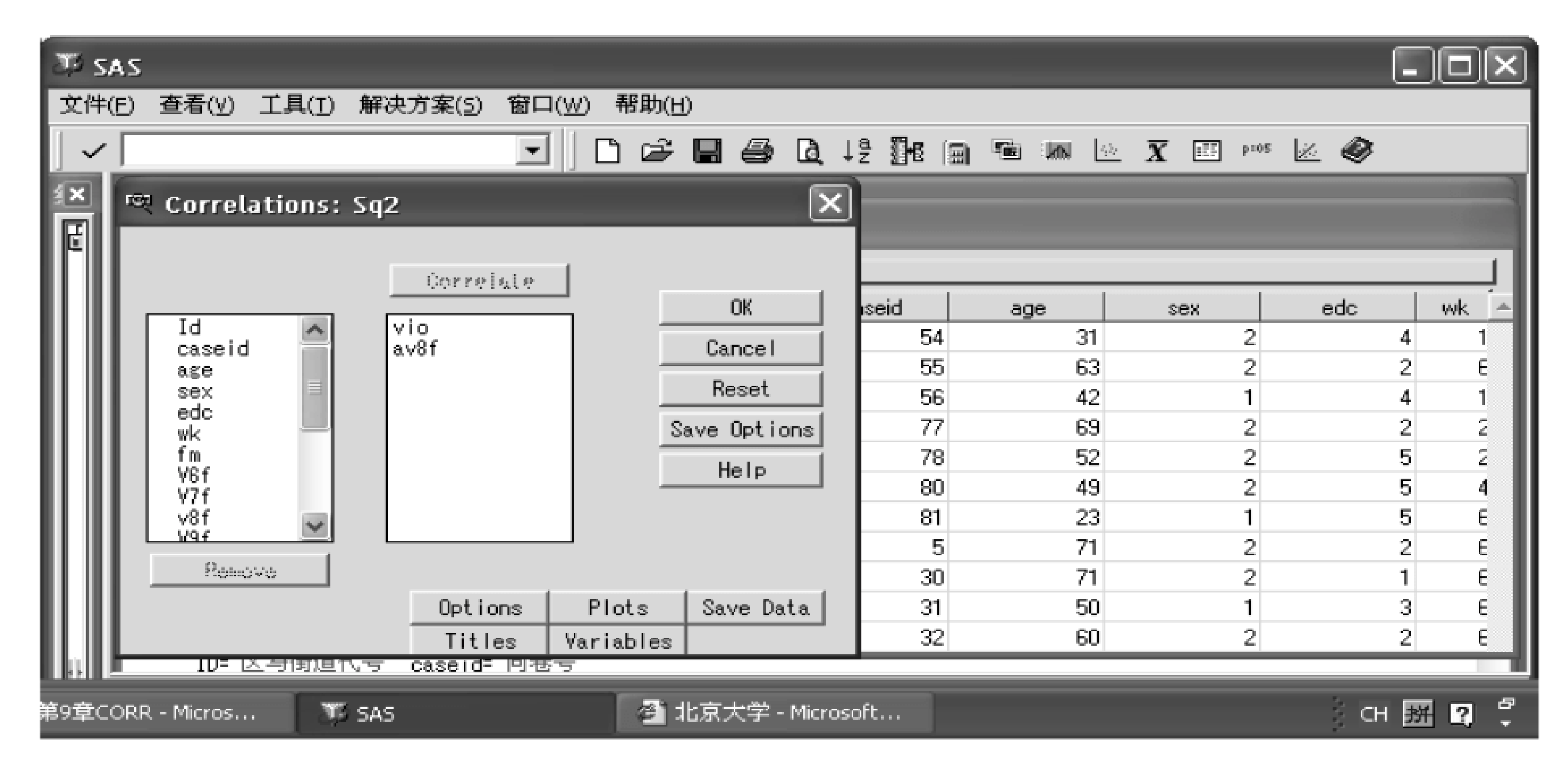


图 9.14 选择人均月结余与人均居住面积两个变量

(7) 单击 Options,进入图 9.15 选择 Kendall's tau-b 相关系数。

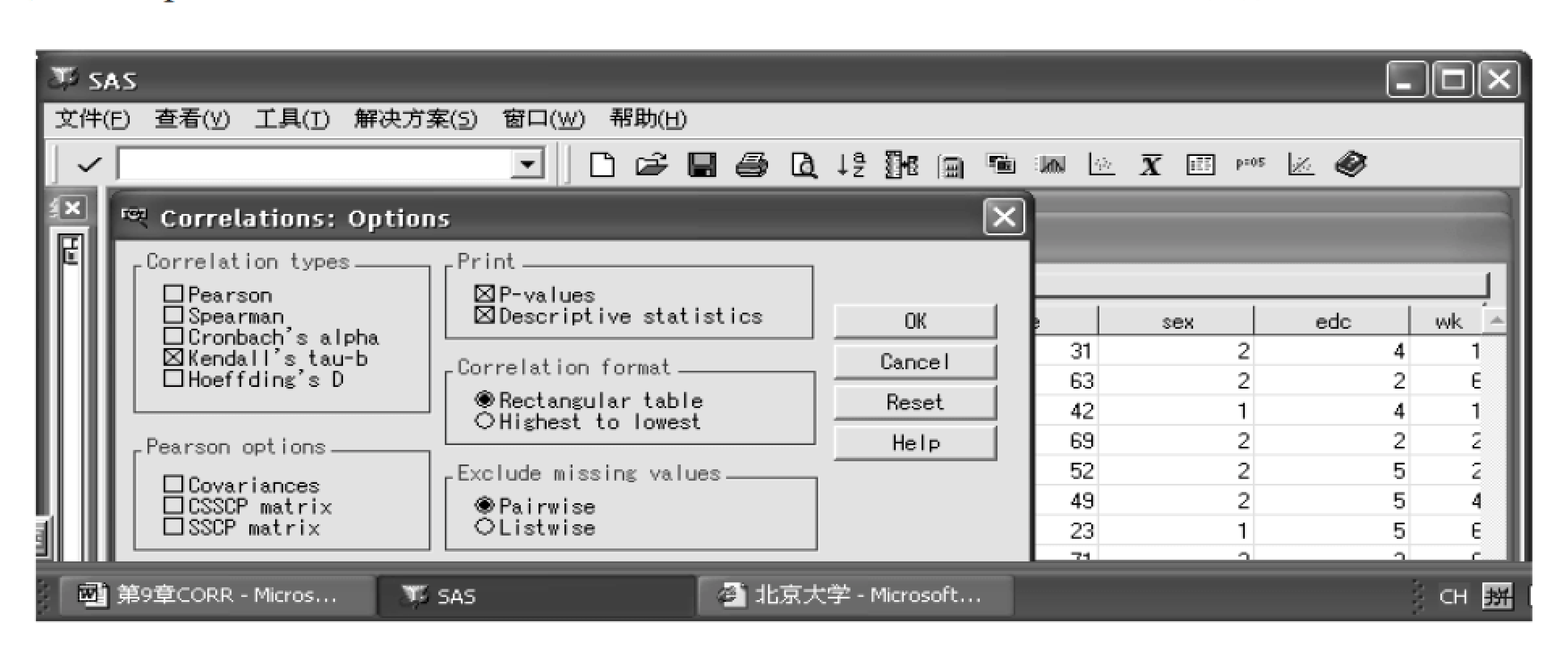


图 9.15 选择 Kendall's tau-b 相关系数

(8) 单击两次 OK 按钮,输出图 9.16 所示的结果。

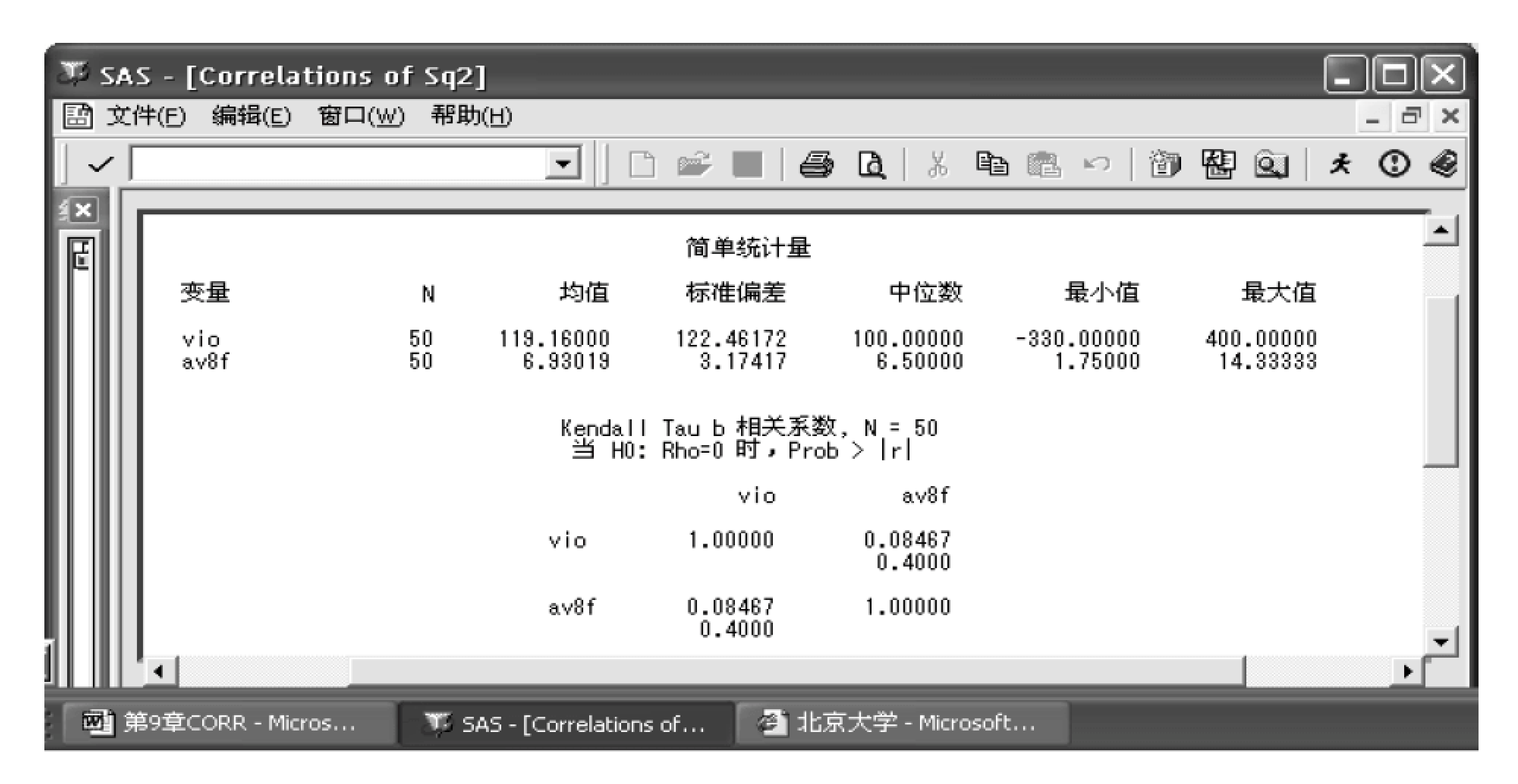


图 9.16 肯氏相关系数 τ_b 的结果

9.4.3 肯氏相关系数 τ_b 结果分析

如图 9.16 所示:

概率值 P=0.4 很不显著,显然不可以拒绝相关系数是 0 的假设,而且样本的相关系数只有 0.08467,接近于 0,说明从此样本看,人均月结余 Vio(等级定序数据)与人均居住面积 Av8f(等级定序数据)之间几乎不存在肯氏等级相关。

9.5 计算次序—比率数据的肯氏相关系数

所谓次序—比率数据是指只有一个变量是次序(或等级)数据,另一个变量是定距以上的百分比数据。

9.5.1 次序—比率数据例子

上一节是计算等级数据的肯氏等级相关系数,本节试计算非等级数据的肯氏相关系数。数据见表 9.1。

表 9.1 文化程度与消费比率变量的关系示例

文化程度(定序变量 edc)	月收入(元)	月支出(元)	消费比率=月支出/月收入(百分数变量 oi:单位%)
1. 文盲	400	300	75%
2. 小学	500	400	80%
3. 初中	600	500	83%
4. 高中(含中专)	700	600	86%
5. 大专以上	800	700	88%

9.5.2 计算次序—比率数据的 Eta 系数

计算“次序—比率”数据应该用 Eta 系数，但是 SAS 系统暂缺 Eta 系数，笔者只好用肯氏相关系数作为参考(注：SPSS 中的 CROSSTABS 过程有 Eta 系数)。

(1) 数据：见图 9.17 所示的社区调查数据。

程序 9.3：见图 9.17。

SAS - [程序编辑器 - 程序9.3.sas]

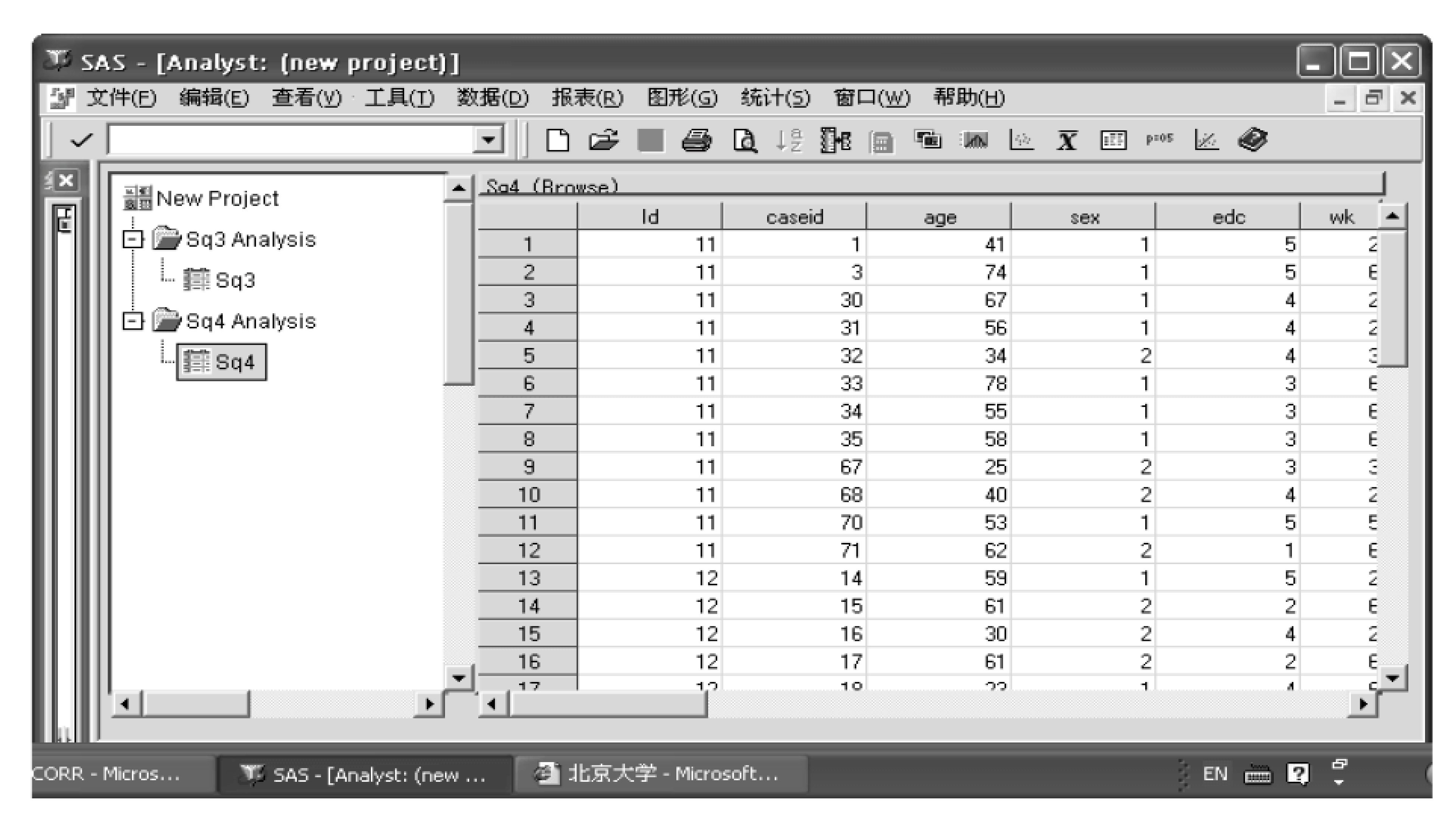
```
DATA sq4;
INPUT  Id 1-2 caseid 3-5 age 6-7 sex 8
       edc 9 wk 10 fm 11 V6f 12 V7f 13-14
       v8f  15-16 V9f 17  v10a 18  v10b 19  v10c 20  v10d 21 vi 22-24 vo 25-27;
IF age=0|sex=0|fm=0|edc=0|wk=0|V6F=0|v7f=0|v8f=0|V9F=0 THEN DELETE;
IF vi=0|vo=0|v7f=0|v8f=0 then delete;
   oi=vo/vi; vio=vi-vo; av8f=v8f/v7f;av=vi/v7f;
LABEL   sex='户主性别'  edc='文化程度' wk='具体工作'
      ID='区与街道代号' caseid='问卷号'
      fm='婚姻状况' v6f='几代人?' V7f='总人数' V8f='居住面积' V9f='住房类型'
       v10a= '煤气'  v10b='卫生间'  v10c='暖气'  v10d='自来水'
      Vi='月收入:元' Vo='月开支: 元';

CARDS;
140735721133062031001800400100000080000000000103002335323120000000111050
140744515122031020001350300000008030001010101020033555331200000000001030
RUN;
PROC FORMAT;
VALUE sex 1='1' 2='0';
VALUE  sex 1='男' 2='女';
VALUE edc 1='文盲' 2='小学' 3='初中'
         4='高中或中专' 5='大专以上' ;
VALUE  id 11='东城区东四居委会' 12='东城区安德里居委会'
        13='东城区赵家楼居委会' 14='东城区南小街居委会'
        31='宣武区牛街街道办' 32='宣武区广内街道办' 33='宣武区天桥街道办'
        41='海淀区中关村七区' 42='北太平庄蓟门里居委会';
 VALUE wk 1='工人' 2='干部 (职员)' 3='商业服务人员'
          4='教员' 5='科研或技术人员' 6='其它';
 VALUE fm 1='未婚' 2='已婚' 3='离婚' 4='丧偶';
 VALUE V9F 1='四合院' 2='普通平房' 3='一般楼房' 4='高层塔楼'
          5='其它活动房等';
VALUE    av8f  LOW-5='1' 5-6='2' 6-7='3' 7-8='4'
           8- HIGH='5';
VALUE    av8f  1='人均 5平米以下'     2='人均 5-6平米以内'
           3='人均 6-7平米以内'   4='人均 7-8平米以内'
           5='人均 8平米以上';
VALUE  vio  LOW--200='0' -200--100='1' -100-0='2'
           0-00='3' 100-200='4' 200-HIGH='5';
VALUE Vio 1='月结余:-200至-100元'  2='月结余:-100至00元'
          3='月结余: 00至100元'   4='月结余:100至200元'    5='月结余:200元以上';
PROC CORR;
    VAR  edc oi;
RUN;
```

第9章CORR - Micros...　SAS - [程序编辑器 - ...　北京大学 - Microsoft...

图 9.17　社区调查数据

(2) 数据集：运行图 9.17 中的程序与数据生成图 9.18 所示的 Work.sq4 数据集。



	Id	caseid	age	sex	edc
1	11	1	41	1	5
2	11	3	74	1	5
3	11	30	67	1	4
4	11	31	56	1	4
5	11	32	34	2	4
6	11	33	78	1	3
7	11	34	55	1	3
8	11	35	58	1	3
9	11	67	25	2	3
10	11	68	40	2	4
11	11	70	53	1	5
12	11	71	62	2	1
13	12	14	59	1	5
14	12	15	61	2	2
15	12	16	30	2	4
16	12	17	61	2	2

图 9.18　数据集中的 oi 等变量

（3）选择 SAS 主菜单中的“解决方案”(Solutions)→“分析”(Analysis)命令，鼠标指针移到图 9.19 所示的“分析家”(Analyst)命令上。



图 9.19 Analyst 的菜单位置

（4）选择“分析家”(Analyst)→“文件”(File)→“按 SAS 名称打开”(Open By SAS Name)→Work 命令和按钮，进入图 9.20。

（5）选择文件名 sq4 后单击 OK 按钮，展示 Work.sq4 数据集的内容，见图 9.18。

（6）选择图 9.18 中的“统计”(Statistics)→“描述性统计”(Descriptives)命令，进入图 9.21 并设置变量：edc(文化程度)和 oi(消费比率)两个变量，见图 9.21。



图 9.20 Work.sq2 数据集文件

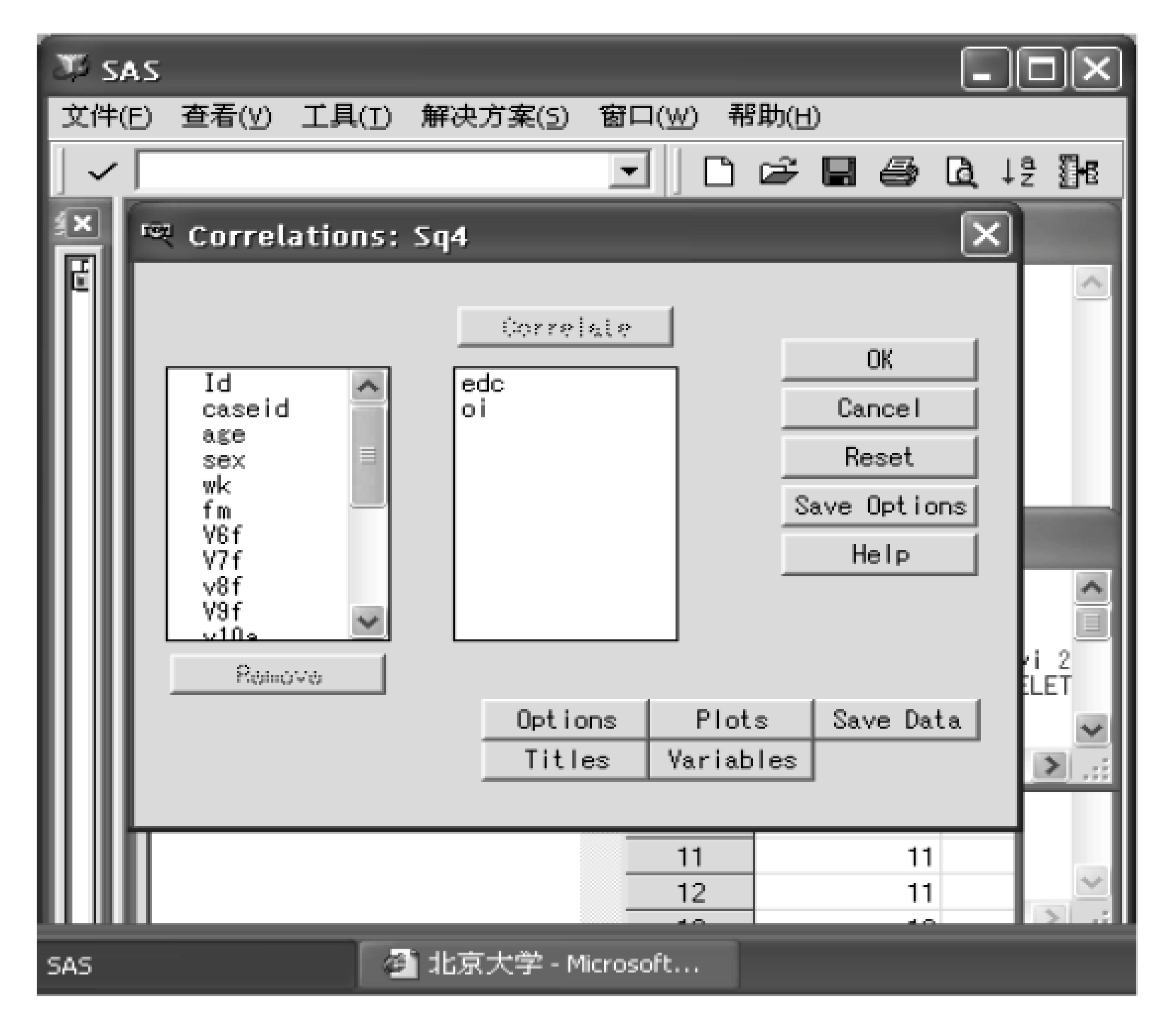


图 9.21 设置 edc(文化程度)和 oi(消费比率)两个变量

（7）单击 Options 按钮选择 Kendall's tau-b 系数，见图 9.22。

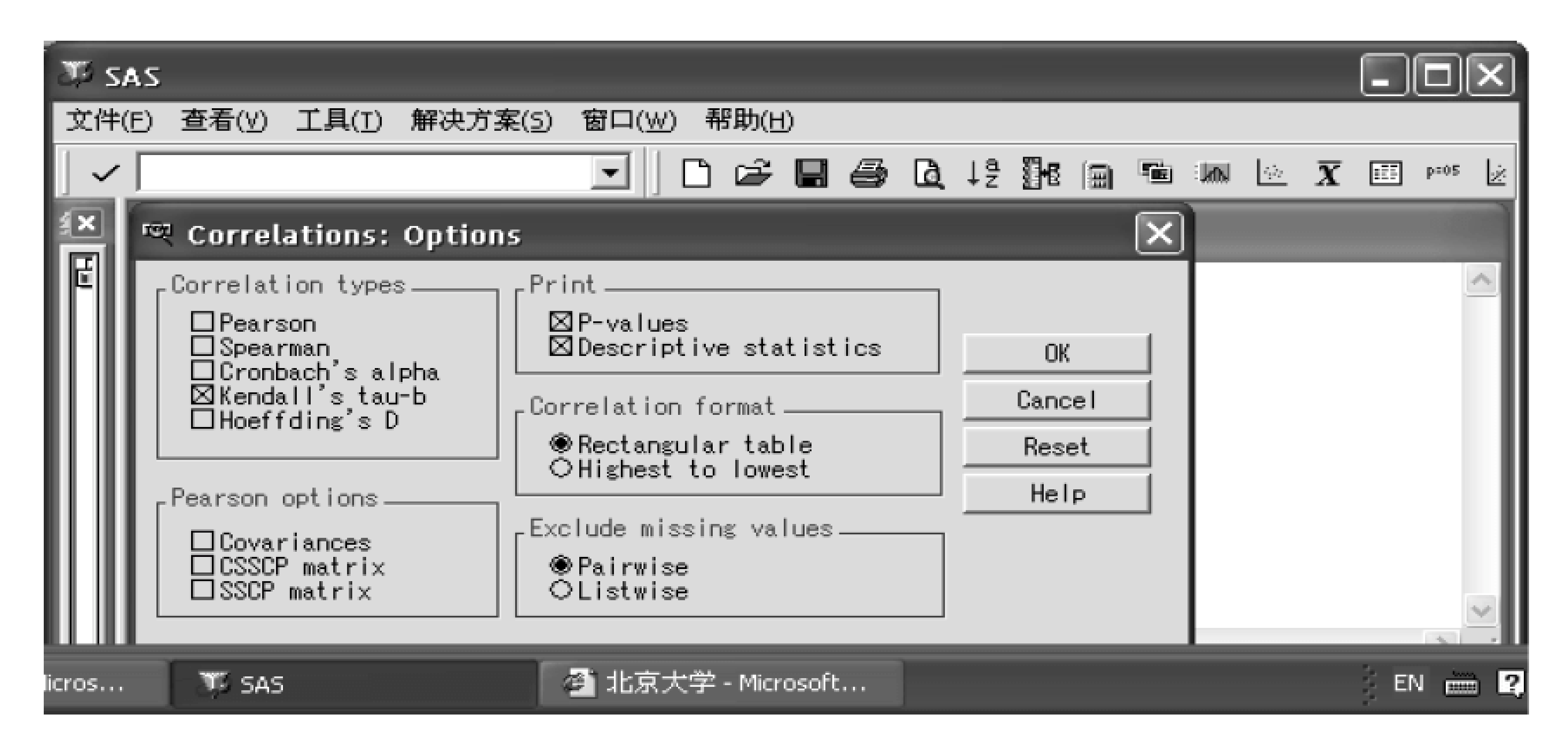


图 9.22 选择 Kendall's tau-b 系数

（8）单击两次 OK 按钮，输出 Kendall's tau-b 系数（如图 9.23）。

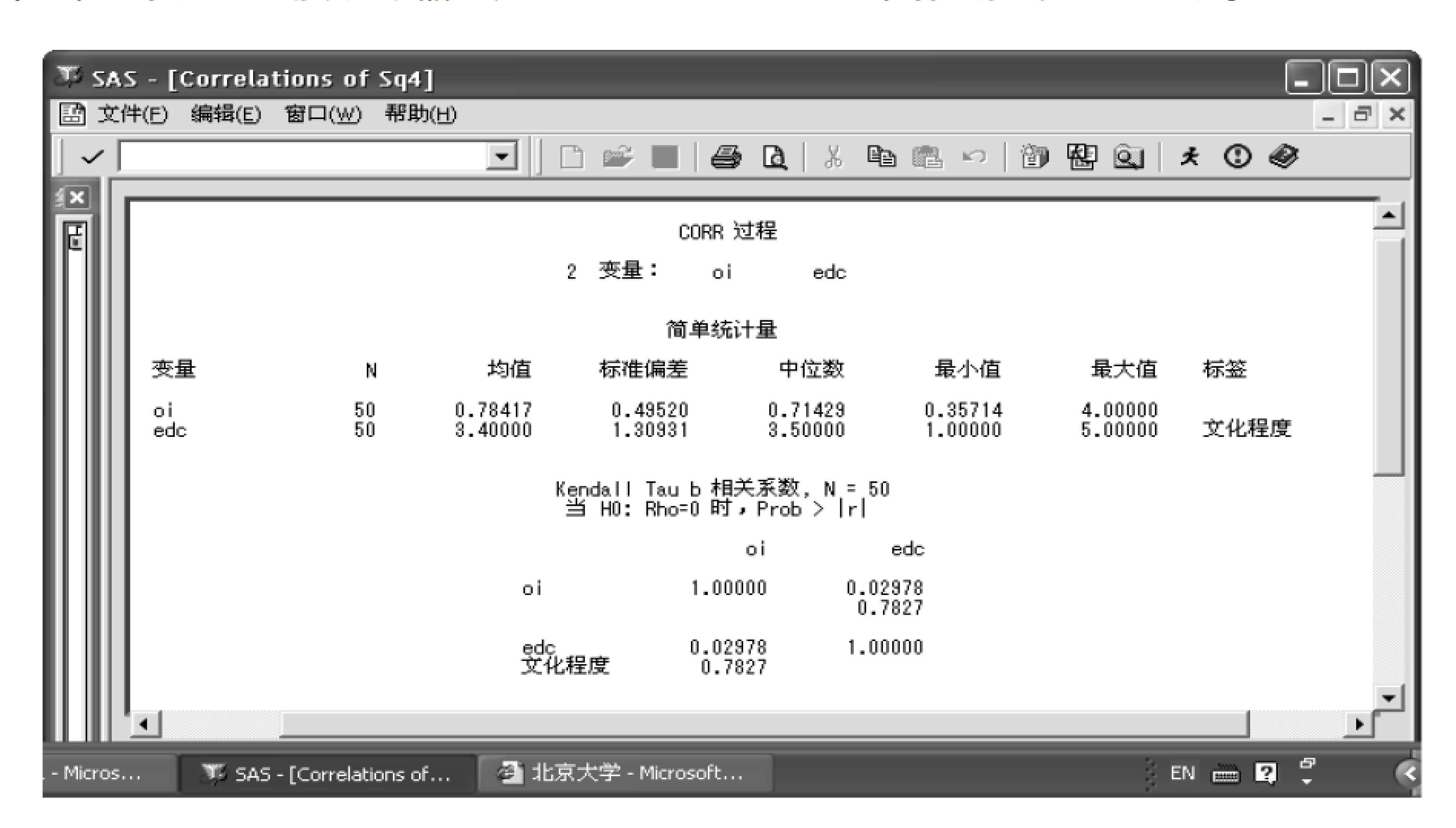


图 9.23 输出 Kendall's tau-b 系数

9.5.3 肯氏相关系数 τ_b结果分析

如图 9.23 所示：Prob > |r|概率值 P=0.7827 很不显著，所以没有理由拒绝相关系数是 0 的假设。而且样本的相关系数只有 0.02978，接近于 0，说明从此样本看，edc（文化程度）和 oi（消费比率）两个变量之间几乎不存在肯氏相关。

9.6 斯皮尔曼等级相关

如果两个变量的数值存在着有序的等级或水平（Level），例如，丈夫的月收入与妻子月收入分别可以划分为：1（2000 元以下），2（2001～3000 元），3（3001～4000 元），4（4001

以上)一共 4 组。则认为丈夫与妻子的月收入分为 4 个等级。在这种情形下,可以采用斯皮尔曼的等级相关公式,计算该对变量是否相关,相关系数有多大。当然,也可以按第 8 章介绍的采用 PROC RANK 过程分析。

9.6.1 斯皮尔曼等级相关系数的计算公式

假设丈夫的变量值为 x,丈夫的变量值平均为 $\overline{X}$。妻子的变量值为 y,妻子的变量值平均为 $\overline{Y}$,那么,利用斯皮尔曼等级相关公式则可求出相关系数 R:

$$R = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum (X_i - \overline{Y})^2 \cdot \sum (Y_i - \overline{Y})^2}} \tag{9.3}$$

公式(9.3)中,R 为斯皮尔曼等级相关系数。

X_i:第 i 个 X 值的等级。

$\overline{X}$:X 的均值。

Y_i:第 i 个 Y 值的等级。

$\overline{Y}$:Y 的均值。

式(9.3)中,分子为双测量的协方差,分母为各变量标准偏差的乘积。

R 的取值范围为:$-1 \leqslant R \leqslant +1$

可见,相关系数与协方差一样,都是测量变量间的线性相关的程度;所不同的是,相关系数是标准化了的协方差。

9.6.2 用“分析家”对话框测量等级相关

下面程序 9.4 中的变量及其数据是北京大学郭崇德教授 1991 年对某些大中城市社区服务研究中的《退休职工调查》数据。笔者从中随机抽取 66 个个案进行分析,其中的职业是从工人至技术人员的等级排序的。

计算每位职工退休前的职业与退休后的职业之间的斯皮尔曼等级相关系数。

进入“分析家”(Analyst)的操作步骤如下:

(1) 运行程序 9.4,产生 SAS 数据集 Work.old4。

程序 9.4:

```
DATA old4;
INPUT id1 1-2 caseld 3-5 n 6 sex 7 age 8-9 edc 10
ocu1 11 ocu2 12 sal1 13-15 sal2 16-18 (v1-v5) (5* 1.);
TITLE  '大中城市社区服务研究《居民调查》数据分析之二';
        /* '对 4区 13个居委会的抽样调查' */
LABEL n= '称谓' sex= '老人性别' age= '老人年龄'
 edc= '受教育年限' ocu1= '退休前职业' ocu2= '退休后职业'
 sal1= '退休前月收入:元'
 sal2= '退休后月收入:元' v1= '医疗费报销与否' v2= '生活能力'
 v3= '干家务事' v4= '与小辈关系' v5= '就医困难情况';
IF sex= 0|age= 0|edc= 0 THEN delete;
```

```
CARDS;
11001117742007815123113
```

```
1202111605660000011322
1202122584660000021325
1202261655261200011310
1202252633260990011310
12023117311600005013110
12023226611600005011410
12024616355600020011212
12025316542600000011410
12025426042600000011410
12026116126620025011322
12032527532606014012213
12033117326600017011320
12033227016600000032220
12034516936600021511210
12034617056600005032413
12035117321610000000000
12036226555620015011310
12037616242625025011310
12037525842618613811310
12038229216600000031210
12038626716600000031310
12045116921110008011210
12045226811100000021410
12046229000000000032110
12047517721609728312204
;
```

(2) 选择图 9.13 中的“统计”→“描述性统计”→“相关”命令，进入图 9.24 并设置变量。



图 9.24　设置 ocu1 和 ocu2 变量

(3) 单击 Options 按钮,进入图 9.25 并选择 Spearman 系数。

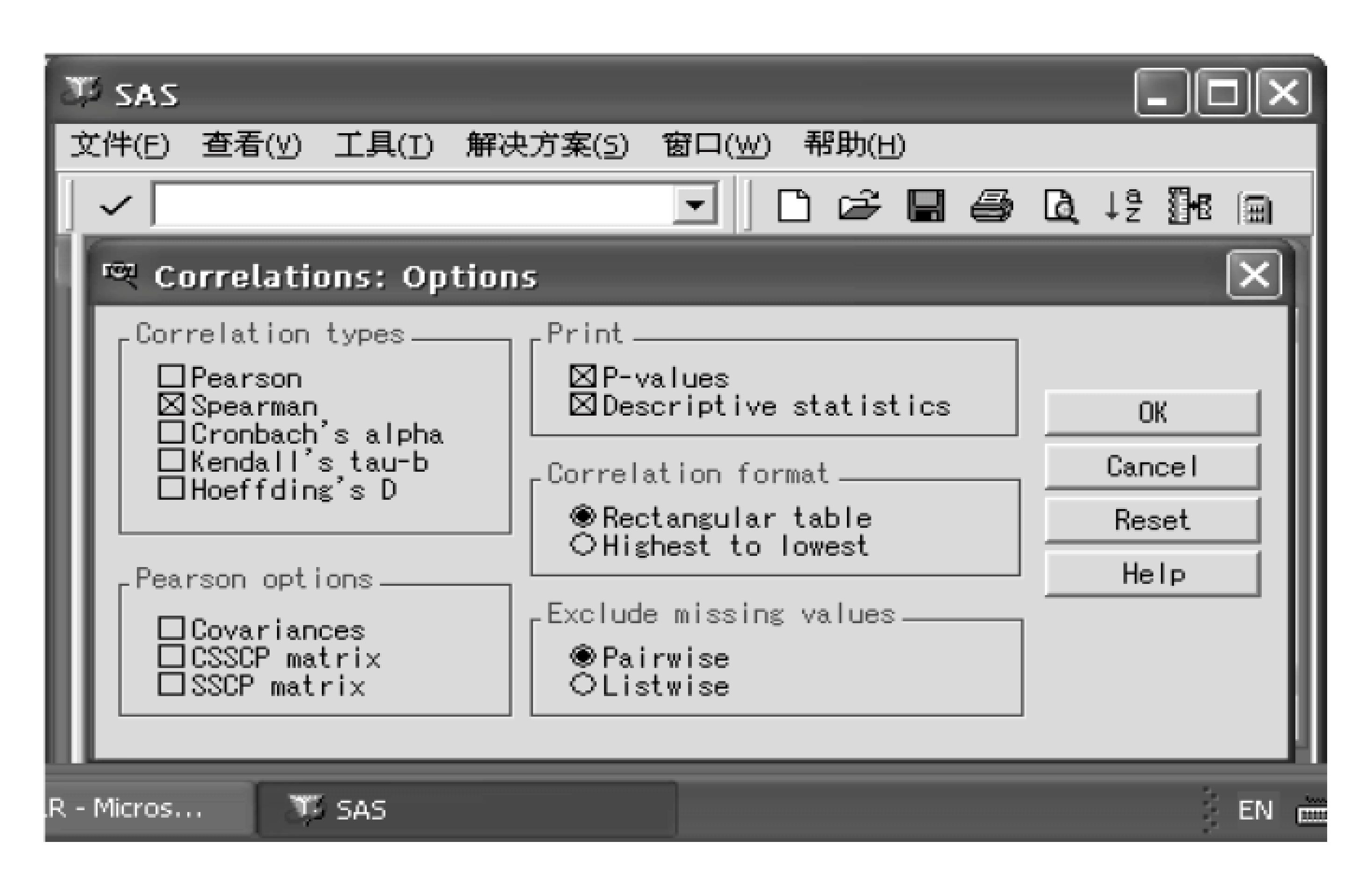


图 9.25　选择 Spearman 系数

(4) 单击两次 OK 按钮,输出图 9.26 所示的结果。

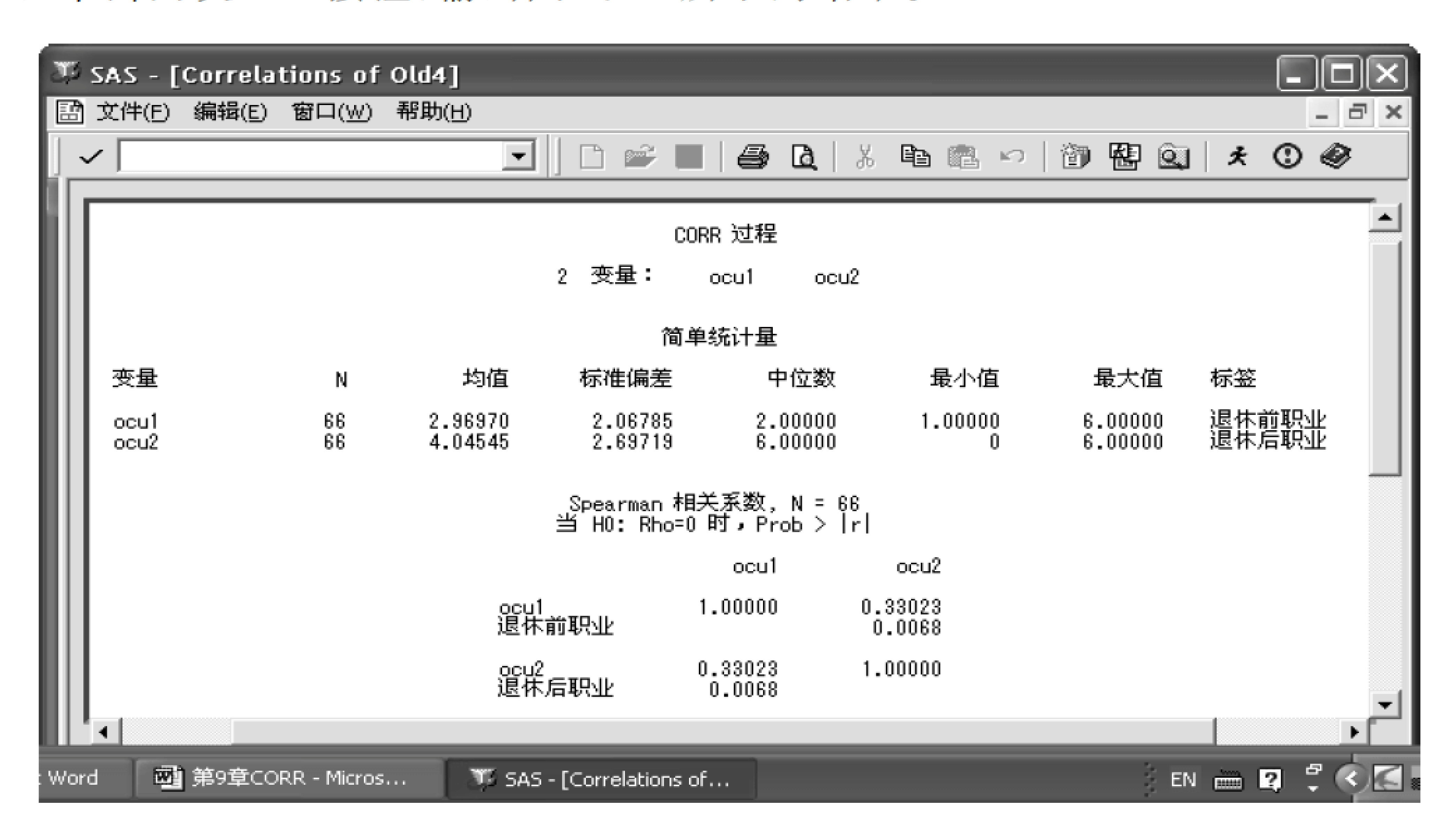


图 9.26　Spearman 相关系数

9.6.3 Spearman 相关系数的分析

如图 9.26 所示:

因为样本的相关系数为 0.33023,而且"当 H0:Rho=0 时,Prob>|r|"的概率值 0.0068<0.05,非常显著,所以样本中每位职工退休前的职业,与退休后的职业之间呈现较强的斯皮尔曼等级相关。

9.7 “标称—标称”型变量的相关测量

Phi(Φ)系数和V系数用于两个变量都是“标称—标称”的相关测量，而V系数是克莱姆(Cramer)的V系数。根据数学公式推导，V系数比Phi系数优化。

程序9.4中的sex与edc两个变量是“标称—次序”型变量，因为没有它们的相关测量，只好把这一对变量降为“标称—标称”型变量进行相关测量，测量的系数用Phi系数和V系数，可用“PROC FREQ; TABLE SEX * EDC/CHISQ;”过程命令获得输出，请参阅图9.27(但无法用对话框命令实现之)。

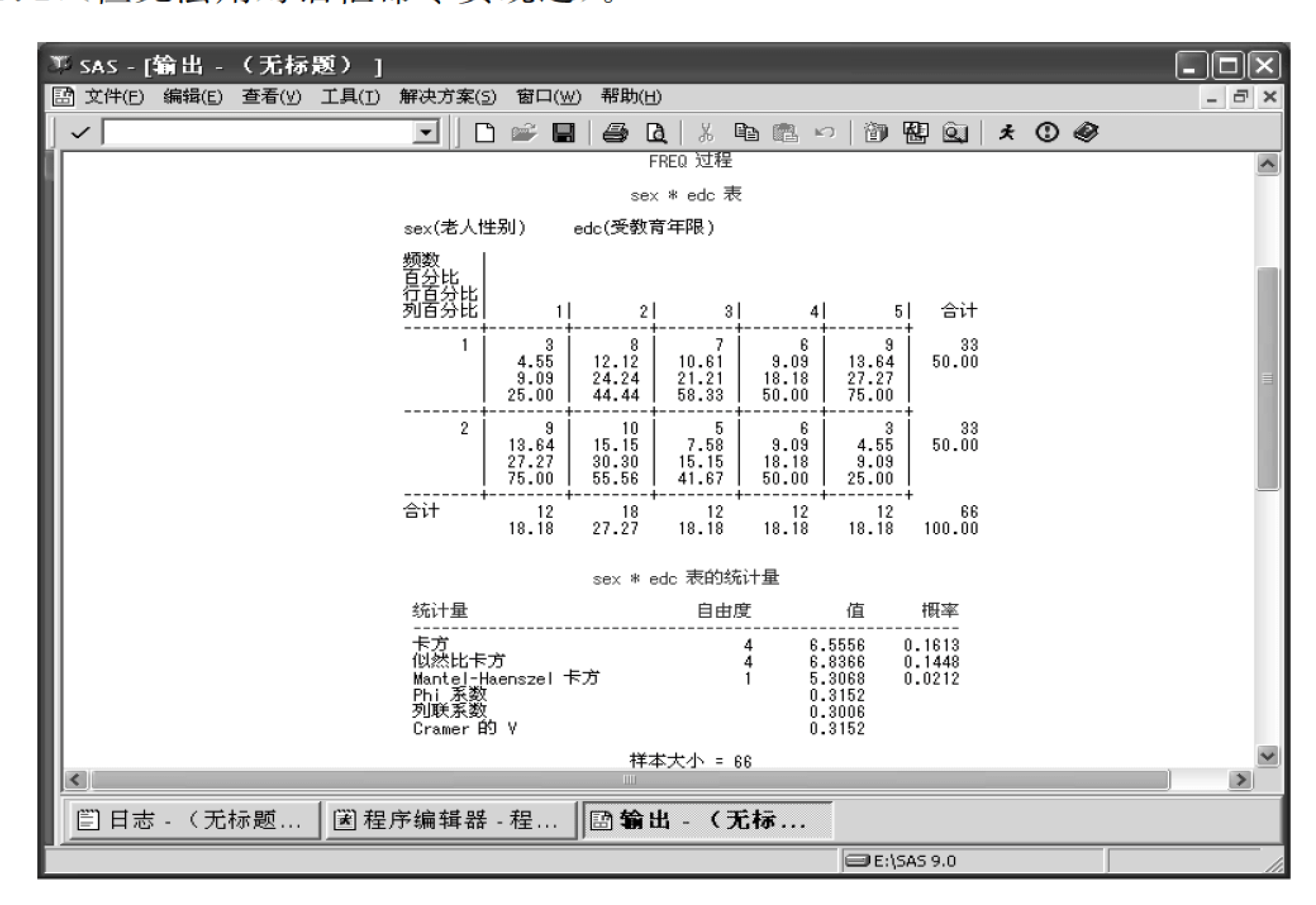


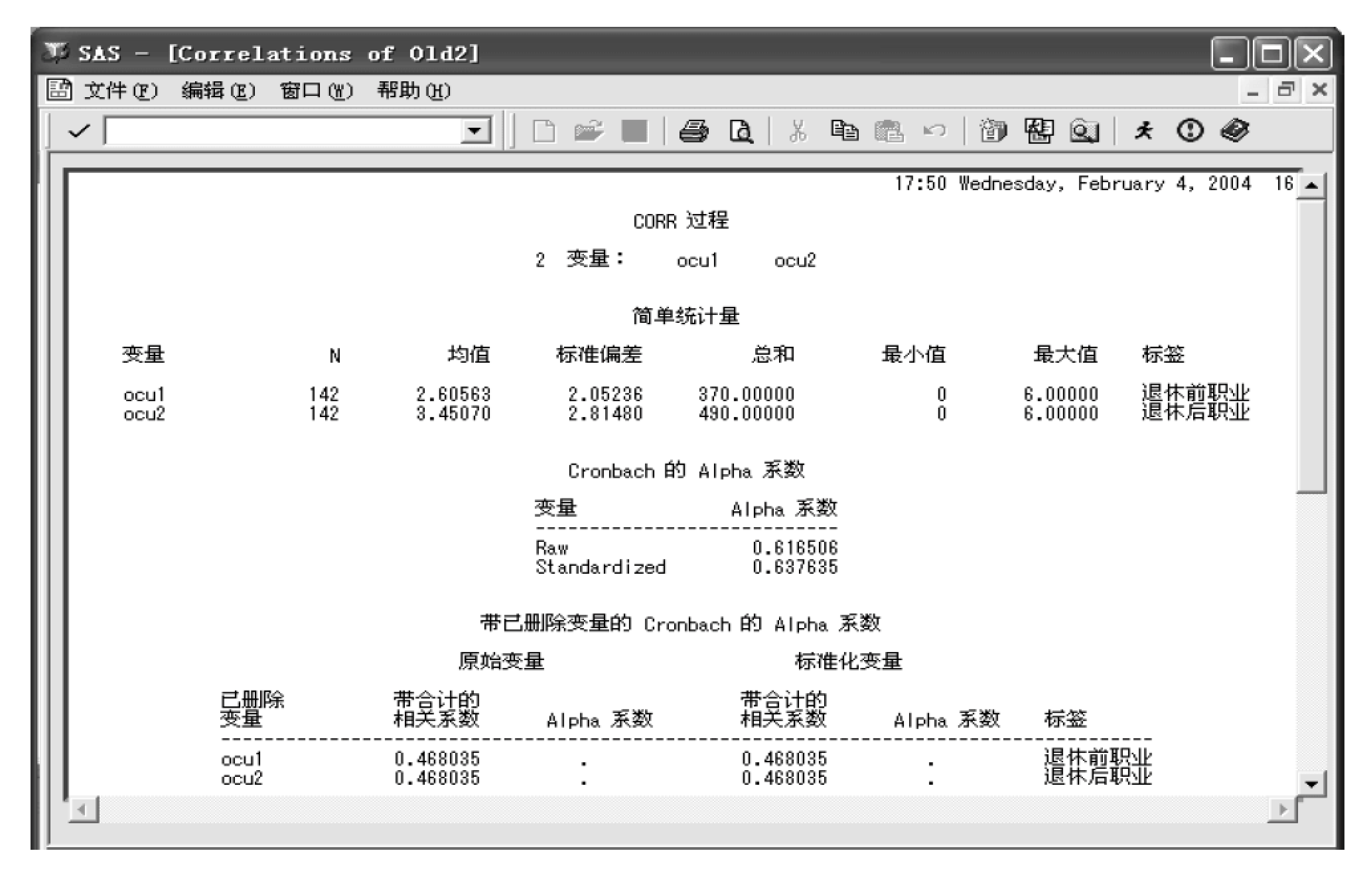
图9.27 “标称—标称”型变量可选用的相关系数

9.8 Cronbach的Alpha系数与Spearman相关系数

图9.28中的Cronbach的Alpha(α)系数给出了一个可靠性系数的下限，该系数的下限等于期望值与实际值之间相关系数的平方。

用程序9.4中的数据计算出退休后职业ocu2与退休前职业ocu1的Cronbach的α理论值如图9.28所示为0.637635，这比Spearman相关系数0.33023大。

注意：Hoeffding的相关系数D可用以测量2个变量具有等级水平时的相关程度。它类似于肯氏(Kendall)等级相关系数τ_b。用法可参阅9.4节。



变量	N	均值	标准偏差	总和	最小值	最大值	标签
ocu1	142	2.60563	2.05236	370.00000	0	6.00000	退休前职业
ocu2	142	3.45070	2.81480	490.00000	0	6.00000	退休后职业

Cronbach 的 Alpha 系数

变量	Alpha 系数
Raw	0.616506
Standardized	0.637635

带已删除变量的 Cronbach 的 Alpha 系数

	原始变量		标准化变量		
已删除变量	带合计的相关系数	Alpha 系数	带合计的相关系数	Alpha 系数	标签
ocu1	0.468035	.	0.468035	.	退休前职业
ocu2	0.468035	.	0.468035	.	退休后职业

图 9.28 Cronbach 的 Alpha 系数

9.9 用 PROC CORR 过程编程计算相关系数

本节介绍编程法，编程法是用 SAS 命令编制程序。编程法往往比对话框法功能齐全。

PROC CORR 过程的主要功能有计算皮尔逊(Pearson)、斯皮尔曼(Spearman)、肯氏(Kendall)以及 Hoeffding 等相关统计量。

1. 尚未加权的情况

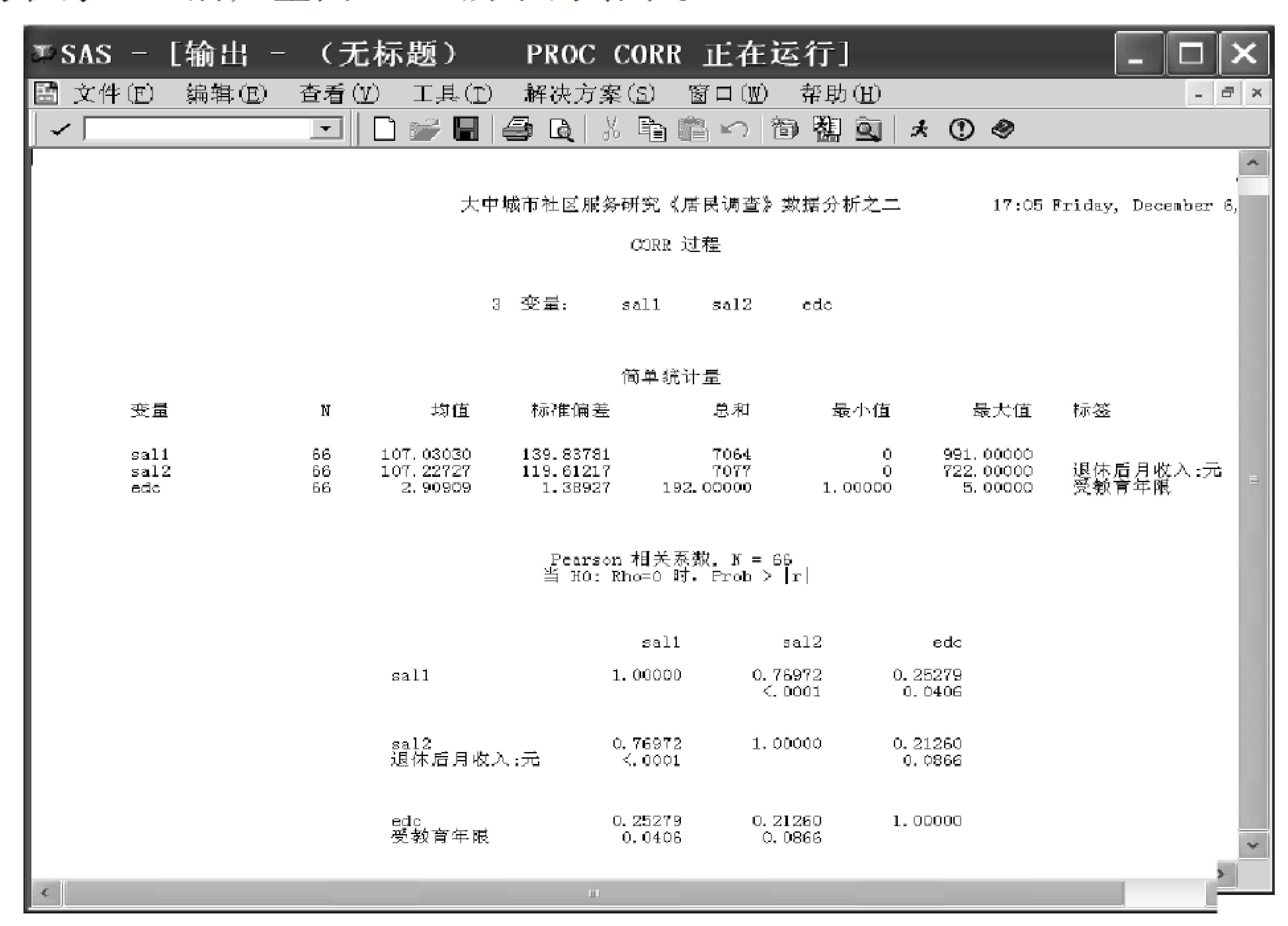
首先介绍没有 WEIGHT 语句时的情形，见程序 9.5 及其输出结果的图 9.29。

程序 9.5：

```
DATA old4;
INPUT id1 1-2 caseld 3-5 n 6 sex 7 age 8-9 edc 10
ocu1 11 ocu2 12 sal1 13-15 sal2 16-18 (v1-v5) (5*1.);
TITLE  '大中城市社区服务研究《居民调查》数据分析之二';
       /* '对 4区 13个居委会的抽样调查'* /
LABEL n='称谓' sex='老人性别' age='老人年龄'
 edc='受教育年限' ocu1='退休前职业' ocu2='退休后职业'
 sal1='退休前月收入:元'
 sal2='退休后月收入:元' v1='医疗费报销与否' v2='生活能力'
 v3='干家务事' v4='与小辈关系' v5='就医困难情况';
IF sex=0|age=0|edc=0 THEN delete;
CARDS;
```

11001117742007815123113
11001717942099172222215
11002526952007809911321
11002617052015017011321
11003517452011018022222
11003626752011516812223
11004226121009007031311
11022616326618015011413
11022525526612009721410
11023227744613010512213
11023427236600000031312
11028616332020000011210
11029518031120020011130
11029627011118018021430
11030516742016019011221
11031515642023000021220
11030625321010509011320
11032116131019017011222
11038228431000011012414
11039316452018000011213
11039426042016000012213
11041525633605611011310
11041616032610018011210
11042527946600000012222
11043517332615010011310
11044526216600000021410
11044616816600000021210
11051626202000000031311
11082516235509010021210
12002616452619522011211
12003525821204006030010
12004517011608506013113
12004626121206405011411
12005617152636036013110
12008428326600000031210
12009116122212531131121
12009225811606007831322
12010525621613012011411
12011425521600000011320
12012516221630017011311
12012626326600000021113
12018226811610008021314
12020626326015020011310
12021116056600000011322
12021225846600000021325
12022616552612000011310
12022526332609900011310
12023117311600005013110
12023226611600005011410

```
120246163556000200112121
202531654260000001141O
```



SAS - [输出 - (无标题) PROC CORR 正在运行]

文件(F) 编辑(E) 查看(V) 工具(T) 解决方案(S) 窗口(W) 帮助(H)

大中城市社区服务研究《居民调查》数据分析之二 17:05 Friday, December 6,

CORR 过程

3 变量: sal1 sal2 edc

简单统计量

变量	N	均值	标准偏差	总和	最小值	最大值	标签
sal1	66	107.03030	139.83781	7064	0	991.00000	
sal2	66	107.22727	119.61217	7077	0	722.00000	退休后月收入:元
edc	66	2.90909	1.38927	192.00000	1.00000	5.00000	受教育年限

Pearson 相关系数, N = 66
当 H0: Rho=0 时, Prob > |r|

	sal1	sal2	edc
sal1	1.00000	0.76972 <.0001	0.25279 0.0406
sal2 退休后月收入:元	0.76972 <.0001	1.00000	0.21260 0.0866
edc 受教育年限	0.25279 0.0406	0.21260 0.0866	1.00000

图 9.29 常用的皮尔逊积差相关系数

从图 9.29 可以看出，未加权的统计结果反映了数据的本来面目，观察值没有被加权。

2. 加权后的情况

现在观察有 WEIGHT 语句时的情形，即程序 9.6 及其输出图 9.30 所示的结果。

程序 9.6：在程序 9.5 的最后增加 weight sex 语句。

```
DATA old4;
INPUT id1 1-2 caseld 3-5 n 6 sex 7 age 8-9 edc 10
ocu1 11 ocu2 12 sal1 13-15 sal2 16-18 (v1-v5) (5*1.);
TITLE  '大中城市社区服务研究《居民调查》数据分析之二';
       /* '对 4区 13个居委会的抽样调查' */
LABEL n='称谓' sex='老人性别' age='老人年龄'
 edc='受教育年限' ocu1='退休前职业' ocu2='退休后职业'
 sal1='退休前月收入:元'
 sal2='退休后月收入:元' v1='医疗费报销与否' v2='生活能力'
 v3='干家务事' v4='与小辈关系' v5='就医困难情况';
IF sex=0|age=0|edc=0 THEN delete;
CARDS;
1100111774200781512311 3
110017179420991722222 15
11002526952007809911321
11002617052015017011321
11003517452011018022222
11003626752011516812223
11004226121009007031311
11022616326618015011413
11022525526612009721410
11023227744613010512213
11023427236600000031312
11028616332020000011210
11029518031120020011130
11029627011118018021430
11030516742016019011221
11031515642023000021220
11030625321010509011320
11032116131019017011222
11038228431000011012414
11039316452018000011213
11039426042016000012213
11041525633605611011310
11041616032610018011210
11042527946600000012222
11043517332615010011310
```

```
11044526216600000021410
```

```
PROC CORR;
VAR sal1 sal2 edc;
WEIGHT sex;      /* 比程序 9.5 增加 weight sex 语句 */
RUN;
```

运行程序 9.6 后产生图 9.30 所示的结果。

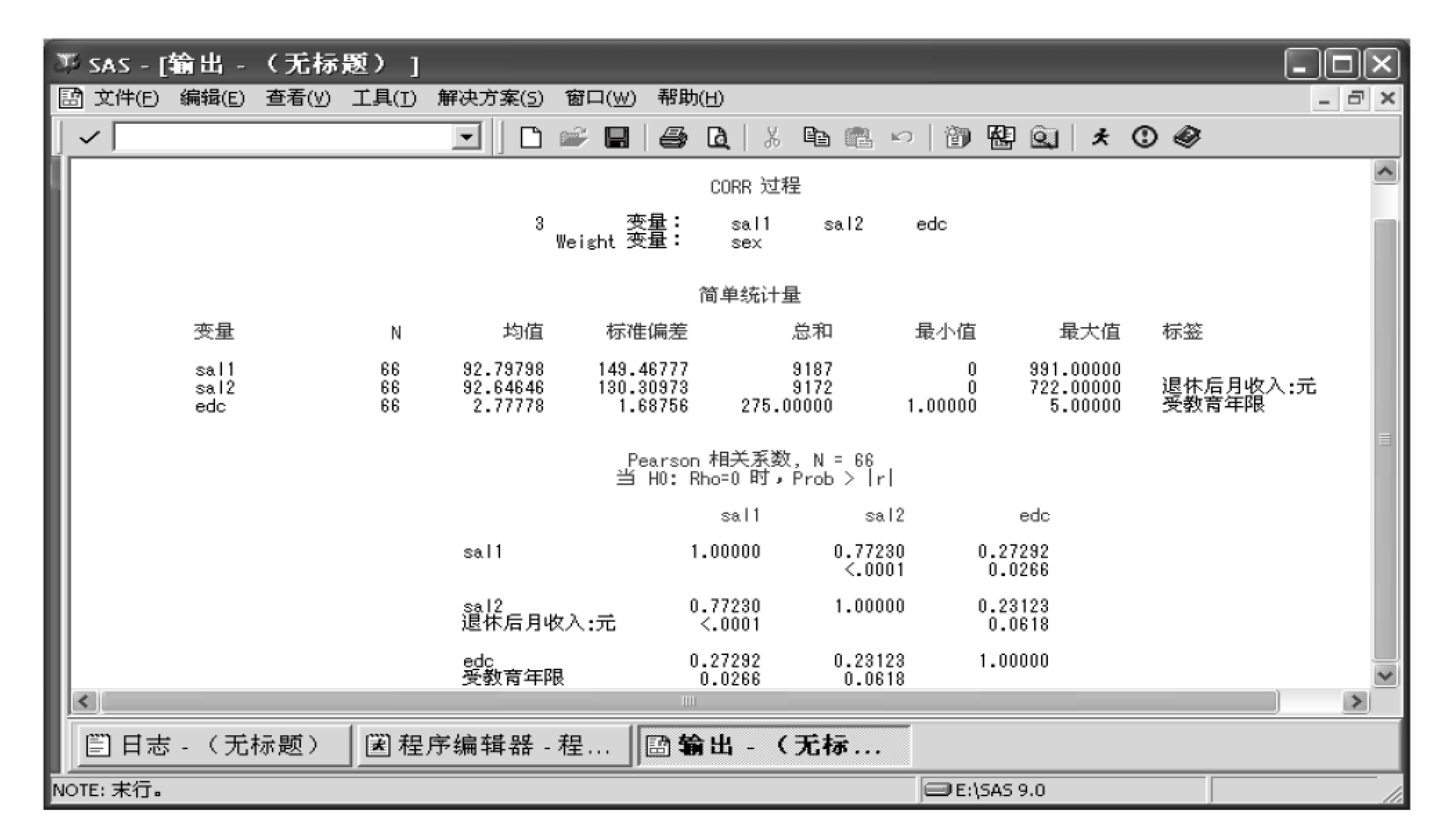


图 9.30 比程序 9.5 增加 Weight 语句后的输出示意图

3. 对图 9.30 统计结果的分析

比较图 9.30 与图 9.29 可以看出,除了 N 栏、最小值栏、最大值栏的数值保持不变外,其他各栏相应的值均被加权了。

所以说,加权与不加权,结果大不一样。如果男女人数相差不太悬殊,或其他变量的各个水平值相差不大时,未必需要加权。

4. 在程序 9.5 中增加 FREQ sex 语句时的前后对比

程序 9.7:在程序 9.5 最后面增加 FREQ sex 语句。

```
DATA sq;
INFILE 'D:\SASDATA\xsq11-12.dat';
INPUT id1 1-2 caseid 3-5 n 6 sex 7 age 8-9 edc 10 ocu1 11 ocu2 12
    sal1 13-15 sal2 16-18 (v1-v5) (5*1.);
TITLE  '大中城市社区服务研究《居民调查》数据分析之二';
PROC CORR;
    VAR sal1 sal2 edc;
FREQ sex;      /* 比程序 9.5增加 FREQ sex 语句 */
```

(1) 运行程序 9.7 后产生图 9.31 所示的结果。

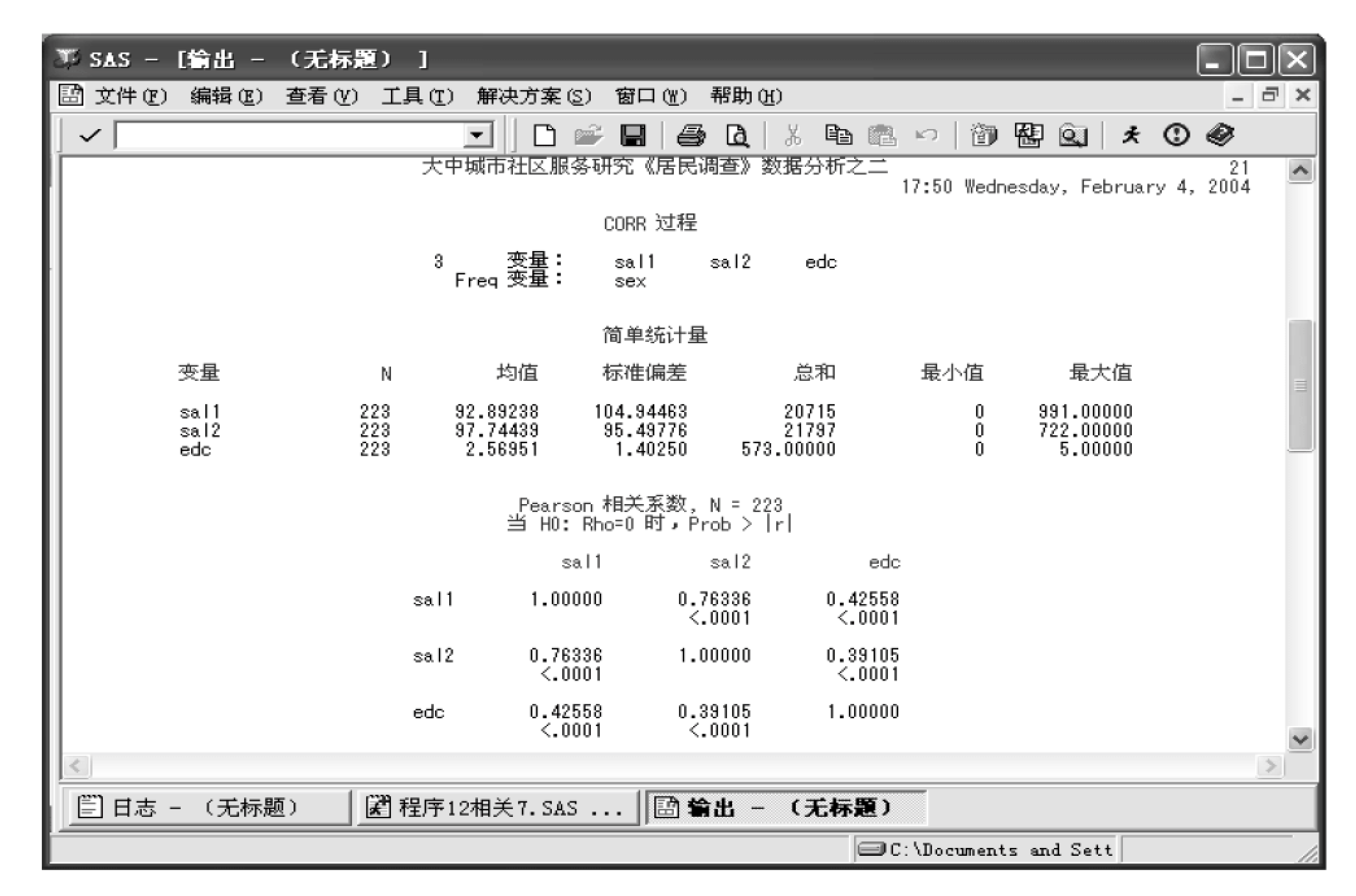


图 9.31 比程序 9.5 增加 Freq 语句后的输出结果

(2) 分析比较：对比图 9.31 与图 9.29 可以看到，由于增加了 FREQ sex 语句，观察值 N 明显地按 sex 值增多到 223(人)。因此图 9.31 比图 9.29 的总和“SUM”相应地增加了。

再比较图 9.31 与图 9.30 可以看出，除了 N 栏的值和“Std Dev 栏”的相应值改变外，图 9.31 其余各栏的值与图 9.30 的一样。

建议：一般情况下应采用程序 9.5。即慎用 WEIGHT 语句和 FREQ 语句。才能保持原来的数据特色。

5. 分组计算双变量的相关程度

下面用“BY 变量表”语句分组计算双变量的相关程度。BY 后面可以指定一个以上的分组变量。

程序 9.8：在程序 9.5 最后面增加“PROC SORT;BY id1;”语句。

```
PROC SORT;BY id1;
PROC CORR;
VAR sal1 sal2 edc;
BY id1;
RUN;
```

运行程序 9.8 后产生图 9.32 和图 9.33 所示的结果。

程序 9.8 说明如下：

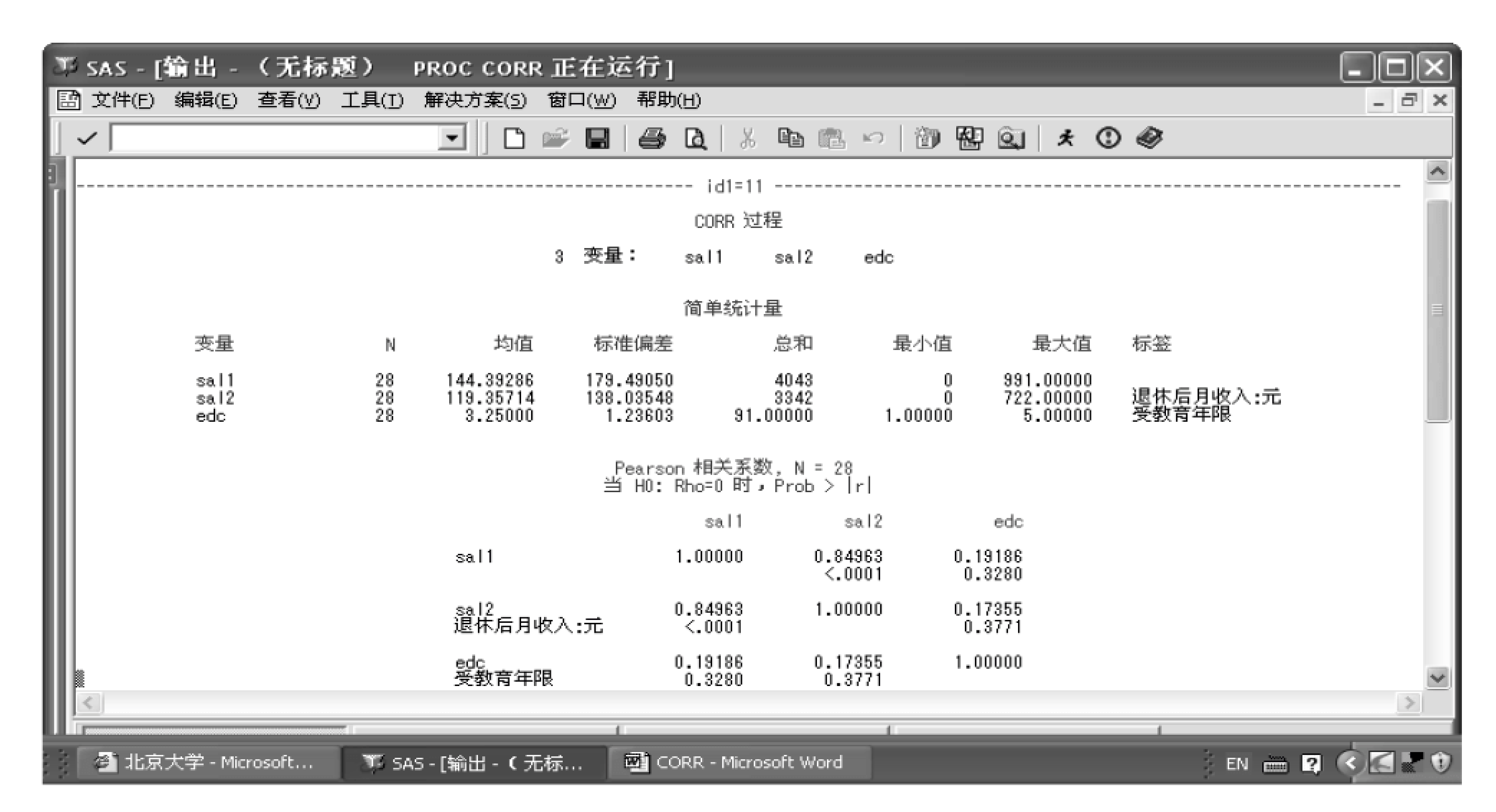


图 9.32 比程序 9.5 增加“BY id1(地区)”语句后的输出示意图

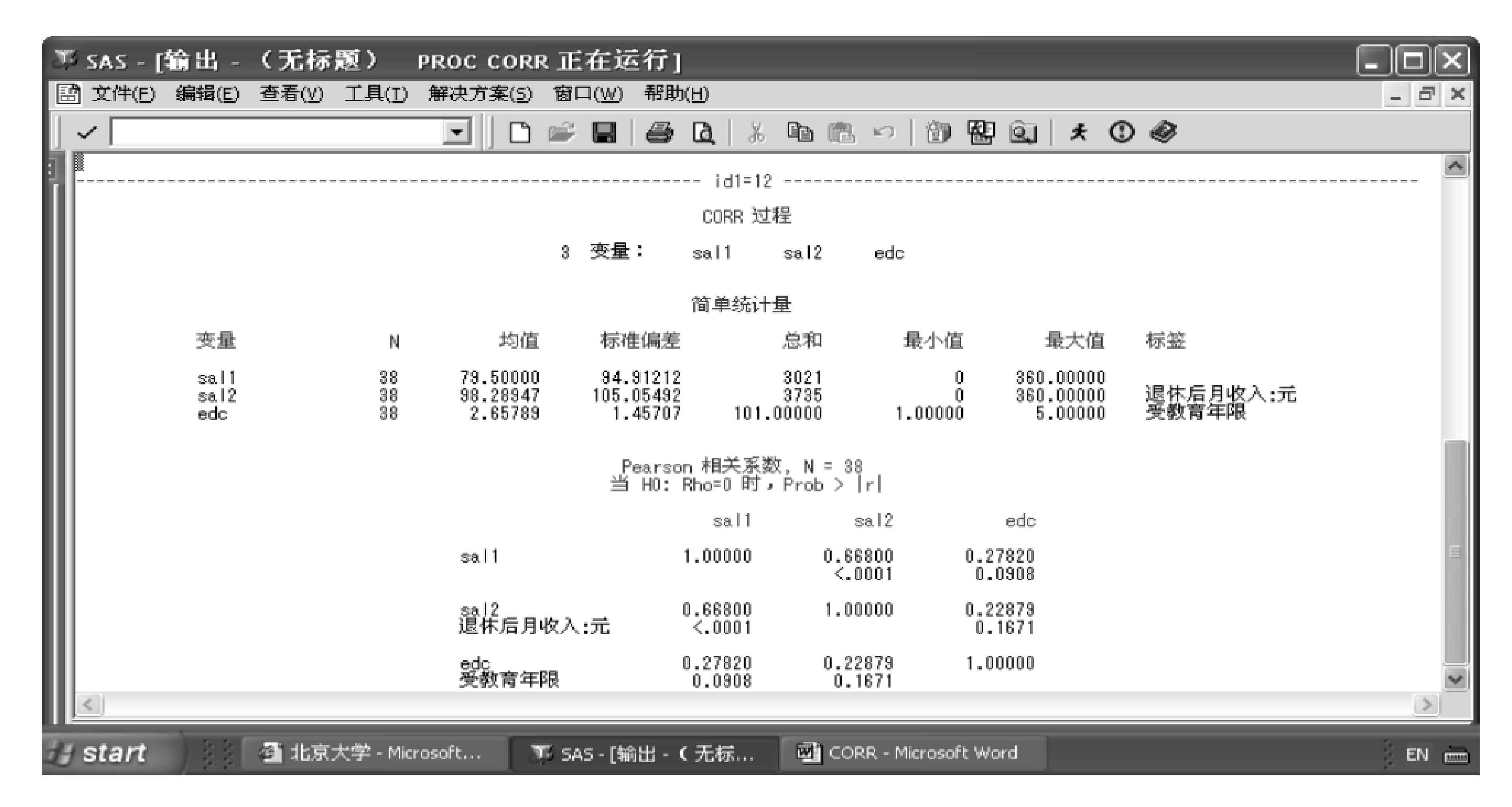


图 9.33 比程序 9.5 增加“BY id1(地区)”语句后的输出(续)

程序 9.8 因为比程序 9.5 增加了一条 BY 语句，所以比图 9.29 增加了一个子图(见图 9.33)，即图 9.29 是不分地区画出的 sal1、sal2 和 edc 三个变量相关分析的总图形。

而且，图 9.32 和图 9.33 则是按照地区为东城区和西城区时，分别画出三个变量的相关分析子图。

6. 对图 9.33 的相关分析

(1) sal2(退休金)与 sal1(原工资)之间的相关系数为 0.6680，且 P 值 $0.0001<\alpha$ 理论值 0.05，所以这一对变量呈现强相关。

(2) sal2(退休金)与 edc(文化程度)之间的相关系数为 0.22879，且 P 值 $0.1671>\alpha$

理论值 0.05，所以这一对变量呈现弱相关但不可靠。

说明：对图 9.32 的分析以此类推。

习 题 9

1. 变量有哪 4 种类型？
2. “比例—比例”（定比—定比）型的变量要用哪一种相关测量？
3. “二分变量—区间以上变量”要用哪一种相关测量？
4. “次序—次序”（定序—定序）型的等级变量要用哪一种相关测量？
5. “次序—次序”型的等级变量还可以用哪一种相关测量？
6. “次序—比率”（定序—定比）型的数据要用哪一种相关测量？
7. “标称—标称”（定类—定类）型的变量要用哪一种相关测量？
8. 哪一个相关系数的值要乘以 30 倍？
9. 试写出计算身高（Height）与体重（Weight）的 Spearman 相关系数。
10. 试分析图 9.34 的结果。

```
                         The CORR Procedure

              3      Variables:    sal1      sal2      edc
               Freq Variable:      sex


                         Simple Statistics

Variable          N         Mean      Std Dev          Sum      Minimum      Maximum

sal1           1221     81.84521     98.10341        99933            0    991.00000
sal2           1221     96.04095     91.24199       117266            0    722.00000
edc            1221      2.33825      1.34204         2855            0      5.00000


                         Pearson, N = 1221
                      Prob > |r| under H0: Rho=0

                         sal1          sal2           edc

             sal1     1.00000       0.67950       0.41820
                                     <.0001        <.0001

             sal2     0.67950       1.00000       0.43077
                       <.0001                      <.0001

             edc      0.41820       0.43077       1.00000
                       <.0001        <.0001
```

图 9.34 皮尔逊积差相关

第 10 章 chapter 10

用 GLM 过程进行回归分析

相关分析是按相关系数的大小来判定两个变量 X_1 与 X_2 之间相关程度的强弱，以及相关趋向的异同，但它不能用以表明 X_1 与 X_2 两个变量的因果关系，即无法测定两个变量的函数关系，更无法确定两个变量函数关系的方程式。

然而，回归分析是通过自变量来预测另一个随机而且是对应的因变量，并采用方程式（例如，$Y=B_0+B_1X_1+\cdots+B_KX_k+E$）来表示因变量与自变量之间的因果关系。

在 SAS 统计分析软件中，设计了回归分析所用的两种应用过程，其一是 GLM，它是 General Linear Model 的英文缩写，可应用于线性和非线性回归模型的分析；其二是 Regression，只能应用于线性回归模型的分析。

GLM 过程处理、分析数据的方法为广义线性模型法，它不仅可处理次序（定序）变量的数据，而且可分析非次序变量的数据，例如：

1. 简单回归（Simple Regression）；
2. 多元（自变量）回归（Multiple Regression）；
3. 方差分析（ANOVA）：尤其适用于非均衡、非对称的数据的方差分析（Unbalanced data）；
4. 加权回归（Weighted Regression）；
5. 偏相关（Partial Correlation）；
6. 多元方差分析（MANOVA）；
7. 多项式回归（Polynomial Regression），或称高次回归；
8. 协方差分析。

GLM 的工作原理，是使用最小平方法（Least square method 即最小二乘法）去研讨一个线性模型。

目前没有 GLM 的对话框，只能采用语句进行统计分析。

10.1 最小平方法的原理

在实际应用中经常要从若干个自变量中预测一个因变量。例如，根据某个学生高三各门的统考成绩，来预测高考成绩以便填报高考志愿时参考。最小平方法（也称最小二

乘法)就是通过估算一组数据的线性关系,进而说明如何利用此法,来预测线性回归模型中的参数。并确定出最佳的回归方程式——模型。

假设线性回归模型为:

$$Y_1 = B_0 + B_1 X_1 + \cdots + B_K X_k + \varepsilon_i \tag{10.1}$$

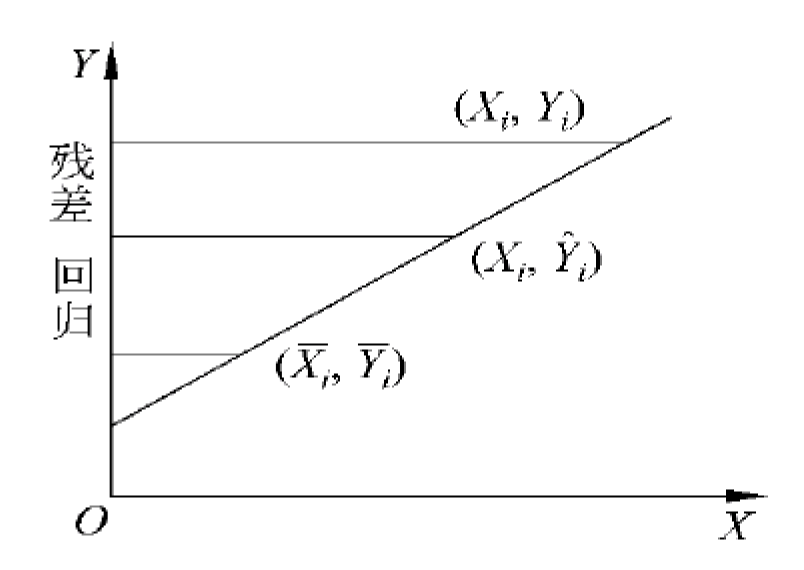


图 10.1 回归分解图

图 10.1 中:

$$\underset{\text{总平方和}}{Y_i - \overline{Y}_i} = \underset{\text{残差平方和}}{(Y_i - \hat{Y}_i)} + \underset{\text{回归平方和}}{(\hat{Y}_i - \overline{Y}_i)} \tag{10.2}$$

配置最佳线性回归模型应具备以下两个条件:

(1) $\sum (Y_i - \hat{Y}_i)^2$ 为最小值;

(2) $\sum (Y_i - \hat{Y}_i)^2 = 0$。

如果所配置的线性方程具备上述两个条件,则采用最小二乘法。

10.1.1 方差分析

将图 10.1 中的总和,以及各个分量分别求平方和之后,可用公式(10.3)表示三者之间的关系:

$$\underbrace{\sum (Y_i - \overline{Y}_i)^2}_{\text{SST}} = \underbrace{\sum (Y_i - \hat{Y}_i)^2}_{\text{SSE}} + \underbrace{\sum (\hat{Y}_i - Y_i)^2}_{\text{SSR}} \tag{10.3}$$

式(10.3)中,SST 为总平方和(Total Sum of Squares),SSR 为回归平方和(Regression Sum of Squares),SSE 为误差平方和(Error Sum of Squares)。

从“总平方和 SST”以及 SSR、SSE 中,可进一步计算出方差的估计量以及非线性方差分析的检验值—F 值,请参阅表 10.1。

表 10.1 回归方差分析表

方差来源	平方和 SS	自由度 df	方差估计量 MS	F 值
回归	SSR	1	MSR=SSR/1	F=MSR/MSE
误差	SSE	$n-2$	MSE=SSE/($n-2$)	
总和	SST	$n-1$		

说明：

- 自由度是指自变量的个数。统计量中每含有一个条件时，则失去一个自由度。
- SSR只有一个自由度。虽然简单直线回归方程式有两项，但因 $\sum$（预测的 $\hat{Y}_i - \overline{Y}_i$）必须为0，故只有一个自由度。
- SSE因为要求 $\sum$（实际的 Y_i − 预测的 $\hat{Y}_i$）＝0，以及 $\sum X$（实际值 Y_i − 预测值 $\hat{Y}_i$）＝0而失去两个自由度，故SSE的自由度为 $n-2$。
- SST有（$n-1$）个自由度。因 $\sum$（实际的 $Y_i - \overline{Y}_i$）＝0而丧失一个自由度。

10.1.2 统计量 F

统计量 F 值用以检验回归系数 B 是否等于0（全0），进而判定回归直线是否有意义。其假设为：

H_0（原假设）：$B=0$

H_1（备择假设）：$B\neq 0$

若总体满足原假设 $B=0$，那么从 $B=0$ 的总体中，如果做无数次样本（其容量为 n）抽样，可以证明，统计量 F 为：

F＝（回归平方和/自由度）÷（误差平方和/自由度）＝（MSR÷MSE）～$F(1,n-2)$（“～”即趋于），即 F 值将服从自由度为$(1,n-2)$的 F 分布。

因此，如果 $F>F\alpha$，则拒绝原假设，说明总体中存在着线性相关，有必要设置回归直线。如果 $F<F\alpha$，则接受原假设，说明没有必要配置回归直线。此处 α 系数是指显著性水平Significance（一般 $\alpha=0.05$ 或 $\alpha=0.01$）。

10.1.3 回归系数B计算法

回归系数 B 即是回归模型中 B_1、B_2 等回归系数的估计值。

如果式（10.1）采用矩阵表示法，则有：

$$\boldsymbol{Y} = B_0 + B_1\boldsymbol{X}_1 + \cdots + B_K\boldsymbol{X}_k + \varepsilon_i$$

通过数学转换之后，最小平方方程式可改写为：

$$(\boldsymbol{X}^1\boldsymbol{X})\hat{B} = \boldsymbol{X}^1\boldsymbol{Y}$$

或

$$\hat{B} = \boldsymbol{X}^1\boldsymbol{Y}(\boldsymbol{X}^1\boldsymbol{X})^{-1} \tag{10.4}$$

$\hat{B}$ 则是所求的回归系数。

说明：

式（10.4）中，$\boldsymbol{X}$ 是原矩阵，$\boldsymbol{X}^1$ 是 $\boldsymbol{X}$ 的转置矩阵，$\boldsymbol{X}^1\boldsymbol{X}$ 的逆矩阵则用$(\boldsymbol{X}^1\boldsymbol{X})^{-1}$表示。

10.1.4 判定系数 R^2

如果因变量 **Y** 与自变量 **X** 关系密切，当 n 个观测点（Xn，Yn）确定后，总平方和 SST 则为定值（见公式（10.3））。这时，回归平方和 SSR 在总平方和 SST 中所占的比例越大，回归模型解释误差的能力就越强。这种能力则称为判定系数，用 R^2 表示。因此，判定系数 R^2 的公式为：

$$R^2 = SSR \div SST = (SST - SSE) \div SST = 1 - (SSE \div SST) \qquad (10.5)$$

因为 $0 \leqslant SSE \leqslant SST$，所以 $0 \leqslant R^2 \leqslant 1$。

判定系数 R^2 有直观的解释意义。例如，当 $R^2 = 0.8$ 时，表示当知道 **Y** 与 **X** 有线性相关时，可以改善预测程度的 80%，换言之，可用 **X** 解释 **Y** 的 80% 误差。若 R^2 越接近 1，**Y** 与 **X** 的关系程度则越高。

10.1.5 残差分析

在式（10.1）中提到 ε，这个 ε 是回归方程中最后一项，一般称之为残差。

从图 10.1 中可知，

$$\varepsilon_i = Y_i - \hat{Y}_i \qquad (10.6)$$

ε_i 应满足 $N(0, \delta^2)$ 的正态分布。根据中央极限定理，计算出的（$\varepsilon_i \div s$）值若趋向于 N(0,1)，即标准正态分布，则可由残差图形检验回归模型是否合适。

当 $\varepsilon_i = 0$ 时，该线性回归模型是标准回归直线。

10.1.6 DW 统计量 D

DW（Durbin-Watson）统计量 D 用来检验回归方程中是否存在自我相关（Auto-Correlation）。统计量 D 可用图 10.2 说明。

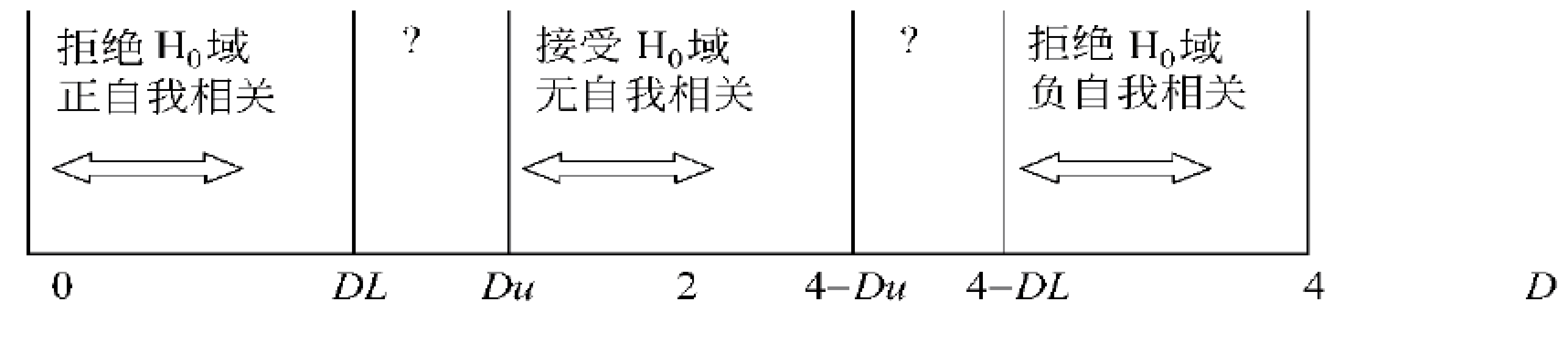


图 10.2 DW 检验的原理图

DW 的假设检验为：

H_0（原假设）：$\rho = 0$（ρ 为总体自相关系数）

H_1（备择假设）：$\rho \neq 0$

- 若 $D < DL$，该区域为正自我相关，则拒绝 H_0 假设。
- 若 $4 - Du < D < 4 - DL$，该区域未有结论。
- 若 $D = 2$，该区域没有自我相关。

• 若 $4-DL<D<4$，该区域为负自我相关。

10.2 GLM中各语句的格式

下面表 10.2 是 GLM 过程的命令语句。

表 10.2 GLM 过程的命令语句

```
PROC GLM [DATA=数据集名称][OUTSTAT=输出的统计量]
         [ORDER=FORMATTED|FREQ|DATA|INTERNAL];
CLASS 分类变量;    /* 此为第二条语句。后面还需 MODEL 配合 */
MODEL 因变量 Y=自变量 X[/NOINT|INT|INTERCEPT|NOUNI|
      SOLUTION|TOLERANCE|E|E1|E2|E3|E4|SS|SS2|SS3|SS4|P|CLM|
      CLI|ALPHA=     |XPX|INVERSE|SINGULAR=1E-8 或某值|
      ZETA=1E-8 或 0];
CONTRAST   '对照说明。10 个汉字,20 个字符' 向量 L 及元素
    [/E|E=effect 或默认为 MS|ETYPE=n|SINGULAR=number];
ESTIMATE   '估计的说明,小于 20 个字符' 值 1 值 2…
    [/E|DIVISOR=number|SINGULAR=number];
LSMEANS effect [/E|E=effect|ETYPE=n|SINGULAR=number|STDERR|PDIFF];
    [MANOVA H= effect E=effect M=equations1,2…
    MNAMES=names PREFIX=name[/PRINTH PRINTE HTYPE=n
      ETYPE=n CANONICAL SUMMARY ORTH]];
    [OUTPUT OUT= 数据集名称 PREDICTED=变量表|P=变量表
    [RESIDUAL=变量表|R=变量表]];
RANDOM effect/Q;
    [REPEATED 因变量名 因子 1 因子 2… (值 1 值 2…)
    [转换的关键词/选择项;]          /* 详见 ANOVA 第 1 章 */
    [TEST H=effects E=effect/HTYPE=ETYPE=;]    /* 详见 ANOVA 第 1 章 */
BY 变量表;
ABSORB 变量表;                  /* 变量表须预先 SORT。此语句使 GLM 无法产生预测值
                                 或输出一个数据集 */
FREQ 变量表;                    /* 按变量表的数值的观察值 n,显示 n 次 */
ID 变量表;                      /* 要求在同一行上显示出各变量值、预测值及残差 */
MEANS effect/选择项;            /* 选择项见 ANOVA 第 1 章的 MEANS 语句 */
```

10.3 GLM程序各语句的使用说明

在 10.2 节中，凡是在中括号[ ]内的命令或关键词(例如，DATA=数据集名称)，均为任选项。带有|符号的任选项视具体需要而定，可任选一项或几项。书写任选项时，符号[ ]或|一并表示分隔符，要用空格替代。

1. 主语句 PROC GLM 的说明

PROC GLM 语句中的 ORDER＝后面的关键词,可选下列 4 个中的某一个:

(1) ORDER＝FREQ:要求按观察值频次降序排序。

(2) ORDER＝DATA :要求按数据输入时的观察值顺序(即原始数据行),显示数据行。

(3) ORDER＝FORMATTED:要求按人为指定的数据格式显示数据行。

(4) ORDER＝INTERNAL:要求按系统默认的格式显示数据行。

2. CLASS 语句

CLASS 语句的一般格式为:

```
CLASS V1 V2 V3…;
```

其中,V1 V2 V3 等变量是分类变量,它们可按需要书写一个或若干个,而且可以是字符型变量或数字变量。如果书写字符型变量,其值限于 10 个字符以内。

注意:如果选用 CLASS 语句,它必须先于 MODEL 语句而位于第二条语句。例如:

```
PROC GLM DATA=OLDER ORDER=FREQ;
    CLASS SEX;
    MODEL NOW=edc salnow* salnow;
```

3. MODEL 语句

MODEL 语句是 GLM 程序中必不可少的语句,用于建立各种回归模型或方差分析模型。

若用 A、B、C 代表分类变量,用 V1 至 V3 代表三个连续变量,可用 MODEL 语句将它们组成如表 10.3 所示的三类模型。

表 10.3　MODEL 语句格式

	MODEL 语句格式	模型的具体名称
回归模型	MODEL Y＝V1	简单一元回归
	MODEL Y＝V1 V2	多元回归
	MODEL Y＝V1 V1 * V2	多项式回归(高次回归)
	MODEL Y1 Y2＝V1 V2	多变量回归
方差分析模型	MODEL Y＝A	单因子 ANOVA
	MODEL Y＝A B C	主效应(effect)模型
	MODEL Y＝A B A * B	含交互效应的因子模型
	MODEL Y＝A B(A) C(B A)	嵌套(Nested effect)模型
	MODEL Y1 Y2＝A B	多变量方差模型(MANOVA)
混合模型	MODEL Y＝A V1	协方差分析模型
	MODEL Y＝A V1(A)	分离斜率(Separate-Slopes)模型
	MODEL Y＝A V1 V1 * A	均一(Homogenity)斜率模型

备注：

```
MODEL Y=A  B  C  A*B  A*C  B*C  A*B*C;
```

等效于

```
MODEL  Y=A|B|C;
```

等号“=”左边为因变量(例如 Y),等号右边为自变量或称独立变量(例如,A B 等)。

下面按照功能类型,逐一列举 10.2 节的表 10.2 中“/”后面的任选项。

(1) 截距选择项

NOINT：表示分析模型中不包含截距参数。

INT(或 INTERCEPT)：要求 GLM 显示出交互效应时截距项的假设检验值(若不指定 INT 则不显示)。

例 10.1：

```
MODEL  Y=V1  V1*V2/NOINT;
```

(2) 结果输出中的选择项

NOUNI：不显示单变量的统计量。

SOLUTION：要求显示标准方程式的解(即：参数的估计值)。

TOLERANCE：要求 GLM 显示 SWEEP(扫描)式子中的容许度 Tolerance。

例 10.2：

```
MODEL Y=V1/SOLUTION TOLERANCE;
```

(3) 对标准假设检验进行控制的选择项

E：要求 GLM 显示所有估计函数的一般格式。

E1：显示每个效应(effect)第一类(TYPEⅠ)的估计函数。

E2：显示每个效应(effect)第二类(TYPEⅡ)的估计函数。

E3：显示每个效应(effect)第三类(TYPEⅢ)的估计函数。

E4：显示每个效应(effect)第四类(TYPEⅣ)的估计函数。

SS1：显示每个效应配合 TYPEⅠ估计函数所产生的平方和 SS。

SS2：显示每个效应配合 TYPEⅡ估计函数所产生的平方和 SS。

SS3：显示每个效应配合 TYPEⅢ估计函数所产生的平方和 SS。

SS4：显示每个效应配合 TYPEⅣ估计函数所产生的平方和 SS。

例 10.3：

```
MODEL  Y=V1/INT  E  E1  SS1;
```

(4) 预测值与残差值的选择项

P：要求 GLM 显示每个观察值、预测值、残差及 DW 统计量。

CLM：显示每个观察值的均值、预测值的置信度。

CLI：显示每个观察值的置信度(Confidence limit)。

ALPHA=P：指定置信区间的 α 值(α=0.01、0.05 或 0.1)。默认为 α=0.05。

例 10.4：

```
MODEL  Y=SEX/SSI  P  CLM;
```

（5）显示中间结果

XPX：要求显示 X'X 矩阵。

INVERSE(或 I)：显示 X'X 矩阵的逆矩阵或一般化矩阵。

例 10.5：

```
MODEL  Y=SEX/CLM  I;
```

（6）调整模型

SINGULAR＝值 n：调整回归模型对线性关系的敏感性。默认值为 n＝1E－8。

ZETA＝值 m：对可估计的 TYPE Ⅲ 与 TYPE Ⅳ 两函数敏感性进行检验。默认为 n＝1E－8。

例 10.6：

```
PROC  GLM;
CLASS  a  b  c;
MODEL  Y=a|b|c/E2  E3  ZETA=1E-6;
```

4. CONTRAST(对照)语句

CONTRAST 语句的一般格式为：

```
CONTRAST  '对照说明'  [向量 L  值 1  值 2…]/选择项;
```

用该语句，可提供一个惯用的、对结果进行假设检验的技巧。例如，可指定一个 $\boldsymbol{L}$ 向量或矩阵，以便检验单变量假设(H_0：LB＝0)或多变量假设(H_0：LBM＝0)。

若假设的条件确实可以检验，在单变量例子中的平方和(H_0：LB＝0)则可按下式计算之：

$$(\boldsymbol{Lb})^1(\boldsymbol{L}(\boldsymbol{X}^1\boldsymbol{X})^{-1}\boldsymbol{L}^1)^{-1}(\boldsymbol{Lb}) \tag{10.7}$$

式中，$\boldsymbol{b}=(\boldsymbol{X}^1\boldsymbol{X})^{-1}\boldsymbol{X}^1\boldsymbol{Y}$。 (平方和 SS 显示在 ANOVA 表上)

CONTRAST 语句中的“对照说明”是一个标签内容，用以说明检验什么内容。一个标签必须对应一个 CONTRAST 语句，而且一个标签的长度必须小于 20 个字符(或 10 个汉字)。标签后面的 L 则是 effect，意指效应。效应分为主效应和交互效应。再后面的值 1、值 2 是具体元素。

例 10.7：在“MODEL Y＝A B;”语句中，假若分类变量 A 有 5 种水平(即：5 个 Level)，分类变量 B 有 3 种值，则 $\boldsymbol{L}$ 向量的元素为：

(μ A1 A2 A3 A4 A5 B1 B2 B3)

H_0：A 合并线性(Pooled A Linear)与 A 二次效应为 0。

为了检验这个原假设，可采用下列 $\boldsymbol{L}$ 矩阵：

$$\boldsymbol{L}=\begin{bmatrix}0 & -2 & -1 & 0 & 1 & 2 & 0 & 0\\ 0 & 2 & -1 & -2 & -1 & 2 & 0 & 0\end{bmatrix}$$

把 **L** 矩阵写成 CONTRAST 语句则是：

```
CONTRAST 'A linear & quadratic'
A  2  1  0  1  2,
A  2  1  2  1  2;
```

CONTRAST 语句中斜杠"/"后面的任选项有：

(1) E：要求显示整个 L 向量。

(2) E=effect：在模型中指定一个 effect 项为误差项。E=effect(效应)若省略不写，则用估计的方差 MS 作为误差项。

(3) ETYPE=n：指明 E=effect 的类型。n=1、2、3 或 4。若指定 E=effect 而不指定 ETYPE=n，分析中则以最高类型来计算。

(4) SINGULAR=number：用于检查估计值。

在对照中的任何一行，若 $ABS(L-L_H)<C*SINGULAR$，则 L 向量无法估计。默认值为 1E-4。其中的 H 是$(X^1X)^{-1}X^1$矩阵。

例 10.8：

```
PROC GLM:
CLASS A B C;
MODEL Y=A|B|C/E2 E3 ZETA=1E-6;
   CONTRAST  "A的线性与二次效应"
            A  -2  -1   0    1  2,
            A   2  -1  -2   -1  2
          /E   SINGULAR=1E-2;
```

5. ESTIMATE 语句

该语句用于估计参数的线性函数，它必须紧跟 MODEL 语句之后。其语句格式为：

```
MODEL …
ESTIMATE '标签内容(≤20个字符)' 向量名称   元素 1   元素 2…
     /E DIVISOR=m SINGULAR=n;
```

其中，m 为除数；n 为估计向量 L 所用的检验值，默认值为 1E-4。(此项同 CONTRAST 中的 SINGULAR 项)。

例 10.9：

```
ESTIMATE '估计 A的线性函数' A 1 -1/E;
```

选择项的说明：

E：要求显示整个 L 向量，见例 10.9 所示。

DIVISOR=m：向量中的各元素除以 m。

例 10.10：

```
ESTIMATE   '…'  A  1  1  -3/DIVISOR=3;
```

等效于

```
ESTIMATE  '…'  A  0.33333  0.33333  -1;
```

SINGULAR=number: number是某一个实际的检验值,默认为1E-4。

6. LSMEANS语句

该语句要求GLM将最小二乘法平均值的计算结果显示在每个效应下,其语句格式为:

```
LSMEANS  A  B  C/E  STDERR  PDIFF
E=effect  ETYPE=n  SINGULAR=number;
```

其中,斜杠"/"后面各任选项说明如下:

E:计算LSMEANS的可估计函数并显示出来。

STDERR:显示LSMEANS估计的标准误差、并计算 H_0:LSM=0的概率。

PDIFF:显示"H_0:LSM(i)=LSM(j)"假设中的所有可能概率值。

E=effect:将某个"effect"当作误差项。若不指定"E=effect",而只指定STDERR或PDIFF,则用均方误差来计算标准误差及概率。反之,若仅指定E=effect而不指定STDERR或PDIFF时,则E=effect将不起作用。

ETYPE=A:见"4. CONTRAST语句"。

SINGULAR=number:同CONTRAST。

例10.11:

```
PROC  GLM;
    CLASS  A  B  C;
    MODEL  Y= A  B  C  A*B;
LSMEANS  A  B  C  A*B;
```

例10.11中的LSMEANS语句要求显示出A、B、C主效位,以及A * B交互效应中每一个水平(Level)的最小二乘法的均值。

7. MANOVA语句

该语句用于计算多元方差分析(Multivariate Anylysis of Variance)。其语句格式为:

```
MANOVA H=effect E=effect M=式 1,式 2,… MNAMES =V1 … ;
```

PREFIX=被转换的变量名/PRINTH PRINTE SHORT CANONICAL SUMMARY;

(本语句中的命令、关键词以及斜杠后面的任选项的用法,请参阅ANOVA有关章节。

8. OUTPUT语句

该语句要求GLM程序产生一个新数据集。预测值(Predicted)、残差值(Residual)以及数据集里的变量值都将存入新数据集里。

OUTPUT语句的一般格式为:

```
OUTPUT OUT=数据集名称 PREDICTED|P=Vn RESIDUAL|R=Vn;
```

说明：PREDICTED|P=Vn 表示"PREDICTED=Vn 或 P=Vn"。其中 Vn 表示 V1,V2,V3,…,Vn 等新变量名。余者类推。新变量名与因变量或 MODEL 中的变量有关。

例 10.12：

```
PROC   GLM;
CLASS   A   B   C;
MODEL   Y=A   B   C   A*B;
OUTPUT   OUT=new1   P=Yhat   R=RESID;
RUN;(注：有些例子没有 RUN。其实微型计算机中的 SAS 可以省略 RUN 语句)
```

例 10.12 说明：

OUTPUT 语句行中，new1 是任意指定的新数据集名称。P=Yhat 为因变量 Y 的预测值，相当于符号 $\hat{Y}$。R=RESID 为 Y 的残差，相当于 $R=Y_i-\hat{Y}_i$。

9. RANDOM 语句

该语句指定模型中具有随机性的 effect(效应)项，以便 GLM 显示出 TYPE Ⅰ～TYPE Ⅳ中每一个 effect 的期望值。其语句格式为：

```
RANDOM   A   B   C/Q;
```

选择项"/Q"，是显示主效应中期望均方的完整的二次形式(Quadratic form)值。

例 10.13：

```
PROC   GLM;
    CLASS   A   B   C;
    MODEL   Y=A   B   C   A*B;
    RANDOM   SEX   EDC/Q;
    OUTPUT   OUT=new1   P=Yhat   R=RESID;
```

10. REPEATED 语句

该语句表示在同一试验单位下，在 MODEL 语句中因变量值具有的重复测量。其语句格式为：

```
REPEATED 因变量名 因子 1 因子 2…  (值 1 值 2…)
  CONTRAST(n)| POLYNOMIAL|HELMERT|MEAN(n)|PROFILE…
    /NOM NOU PRINTM PRINTH PRINTE PRINTRV
      SUMMARY CANONICAL HTYPE=…;
```

该语句的用法详见 ANOVA 有关章节。

11. 其他语句

其他语句如下：

```
ABSORB  V1  V2…;
BY  V1  V2…;
FREQ  V;
ID  V1  V2…;
```

以上 4 条语句参阅本章 10.2 节的相应格式说明。此外，TEST 语句、MEANS 语句详见 ANOVA 的相应语句。

10.4 调用 GLM程序作一元线性回归

通常有这类情形：某公司的经济效益与经营管理模式、员工服务意识、产品的产销对路、售后服务及成本核算等因素有关。农场的收成，与地质的优劣、种子的优劣及田间管理的差异等因素有关。又例如，人口增长率往往与育龄青壮年的婚姻状况、婴儿出生率以及死亡率等要素密切相关；通过这些要素，希望预测某个时期人口增长的情况，以便制定出相应的计划生育措施，诸如此类的问题，应该通过回归分析法解决。

10.4.1 数据与程序

本节引用北京大学郭崇德教授 1991 年关于社区服务的一个调查数据，说明简单线性回归的分析法。为了节省版面，随机抽取北京市东城区 77 位老人（个案）进行分析，分析的变量如下。

EDC：老年人的文化程度

SAL1：退休前的月收入

SAL2：退休后的月收入

V1：医疗保障

V5：就医困难情况等。

下面要求采用 GLM 过程估计退休后的月收入与其文化程度是什么关系，进而建立一个简单线性回归的方程式。

程序 10.1：用 GLM 过程对工资进行简单线性回归分析。

```
DATA Xsql;
INPUT id1 1-2 caseid 3-5 n 6 sex 7 age 8-9 edc 10 ocu1 11 ocu2 12
      sal1 13-15 sal2 16-18 (v1-v5) (5*1.);
LABEL n='称谓' sex='老人性别' age='老人年龄' edc='受教育年限'
     ocu1='退休前职业' ocu2='退休后职业' sal1='退休前月收入：元'
     sal2='退休后月收入：元';
CARDS;
11001117742007815123113
11001717942099172222215
11002526952007809911321
11002617052015017011321
```

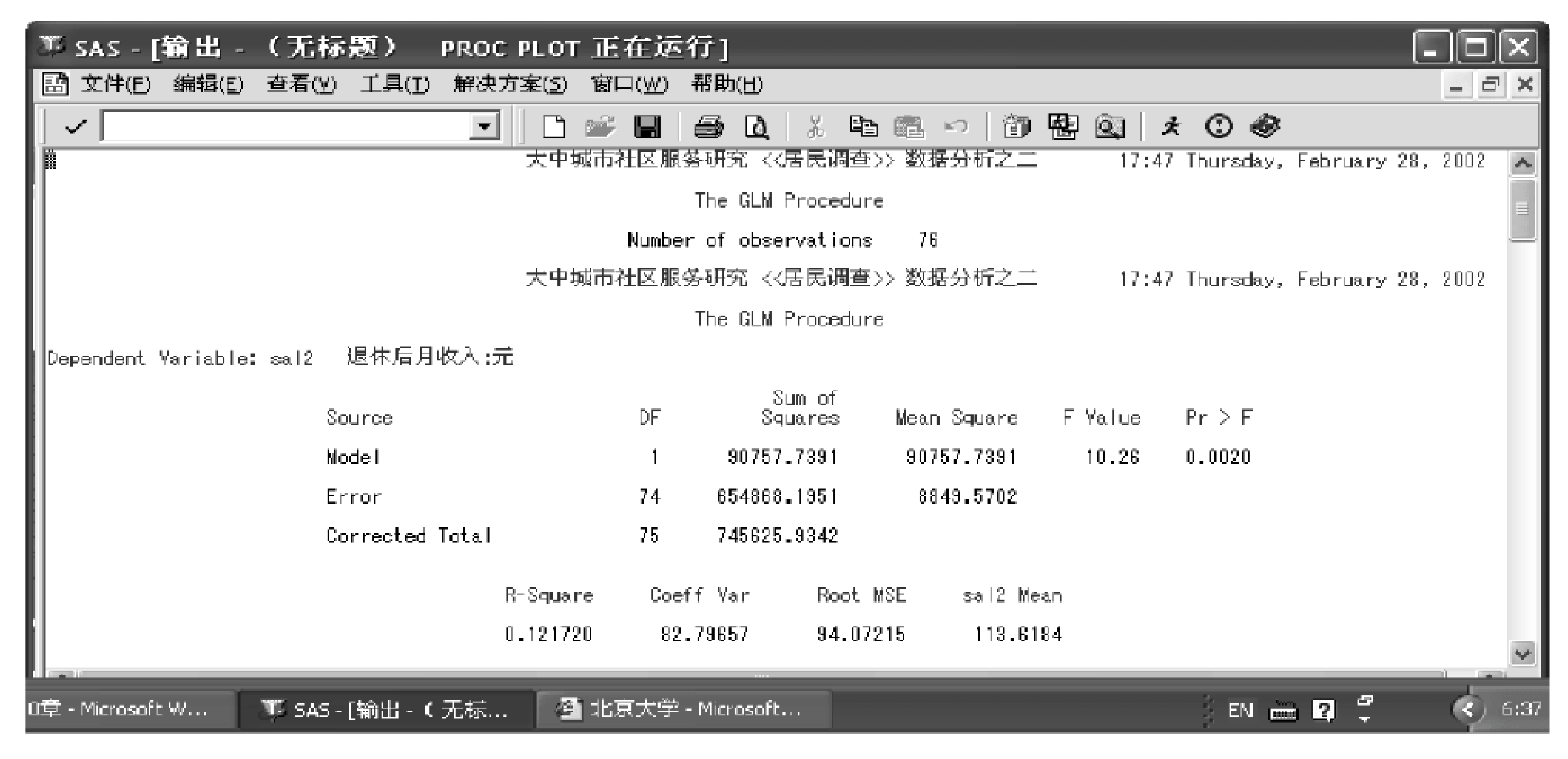
```
110035174520110180222222
110036267520115168122223
110042261210090070313111
110055257326200100113244
110062275210046103213100
110072289100000000311200
110085269211125150112100
110096176230070140121211
110106158422120100113144
110115266400000000313100
110126251426250140214144
110131175216092200112200
110132274116050102113200
110133166220190160114200
110134263210055098112200
110141160320200150114100
110155169520200250114100
110156262320084150114100
110165160420260235112100
110176274200000000314100
110186273460050130114100
110192265226090115213200
110202260116000090214111
110215267520032142214111
110226163266180150114133
110225255266120097214100
110232277446130105122133
110234272366000000313122
110245167400300186112200
110246263200070106213200
110255167540166216114244
110266264211087120214200
110272266166000000314100
110286163320200000112100
110295180311200200111300
110296270111180180214300
110305167420160190112211
110315156420230000212200
110306253210105090113200
110321161310190170112222
110322262200000000313211
110336279100000000323111
110346253420152123121222
110355158311150150112100
```

```
11035625231112012011210
11036526511607012711410
11036616632215015011410
11037526021006114011410
11037617621106013011210
11038118231000011012411
11038228431000011012414
11039316452018000011213
11039426042016000012213
11040627122607513011412
11040517622608615011412
11041525633605611011310
11041616032610018011210
11042527946600000012222
11043517332615010011310
11044526216600000021410
11044616816600000021210
11045617742605013011425
11046728410000000032220
11047727010000000032212
11048525631020014011310
11048615531024000011210
11049116952208015011110
11050516922615013011313
11050626710000000032211
11051516842614023011210
11051626202000000031311
11082516235509010021210
12002616452619522011211
12003525821204006030010
12004517011608506013113
12004626121206405011411
12005617152636036013110
12006516132233033011210
12007116726600000011310
12007226016600000031310
12008428326600000031210
12009116122212531131121
12009225811606007831322
12010525621613012011411
12010615621613513011411
12011315826600000011220
12011425521600000011320
12012516221630017011311
```

```
1201262632660000021113
12013525846623028011210
```

```
12040618052618030012110
12040527754608016021410
12041526316600000021312
12042228200000000021210
12043227500000001321222
12044517333605113811221
12045116921110008011210
12045226811100000021410
12046229000000000032110
12047517721609728312204
;
PROC GLM;
PROC FORMAT;
VALUE n 1="父亲" 2="母亲" 3="岳父(或公公)" 4="岳母(或婆婆)"
        5="户主" 6="爱人" 7="其他(爷爷奶奶等)";
VALUE sex 1="男" 2='女';
VALUE edc 1='文盲' 2='小学' 3='初中'
          4='高中或中专' 5='大专以上';
PROC GLM;
  MODEL SAL2=edc;
  OUTPUT out=out1 p=predict r=residual;
PROC PLOT;
  PLOT sal2*edc predict*edc="p" /OVERLAY;
  PLOT residual*edc/vref=0;
RUN;
```

运行程序10.1产生图10.3至图10.5所示的结果。

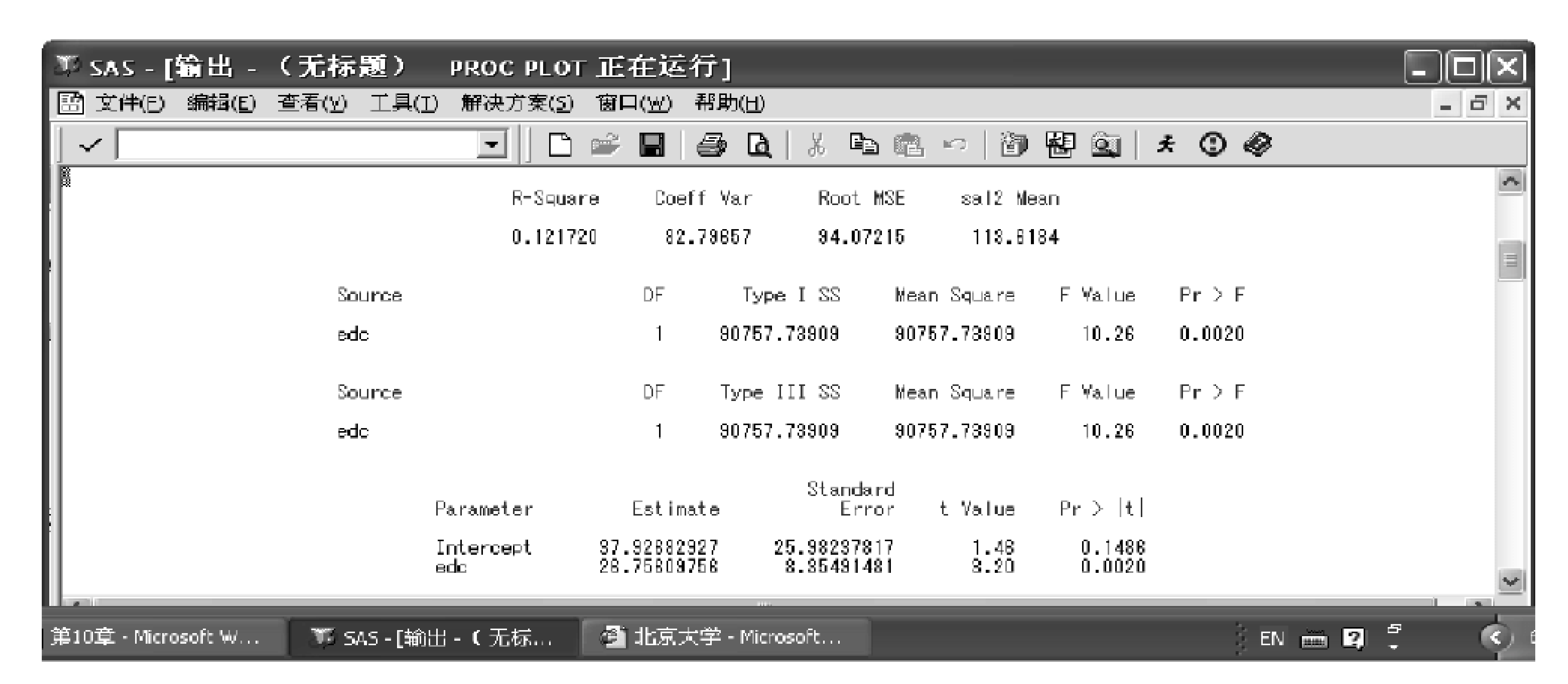


Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	90757.7391	90757.7391	10.26	0.0020
Error	74	654868.1951	8849.5702		
Corrected Total	75	745625.9342			

R-Square	Coeff Var	Root MSE	sal2 Mean
0.121720	82.79657	94.07215	113.6184

(a) 北京市东城区社区服务研究数据分析一

图10.3 简单线性回归输出图


SAS - [输出 - （无标题） PROC PLOT 正在运行]

R-Square	Coeff Var	Root MSE	sal2 Mean
0.121720	82.78657	94.07215	113.6184

Source	DF	Type I SS	Mean Square	F Value	Pr > F
edc	1	90757.73809	90757.73809	10.26	0.0020

Source	DF	Type III SS	Mean Square	F Value	Pr > F
edc	1	90757.73809	90757.73809	10.26	0.0020

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	97.92882927	25.98237817	1.48	0.1486
edc	28.75809758	8.85491481	3.20	0.0020

(b) 北京市东城区社区服务研究数据分析

图 10.3（续）

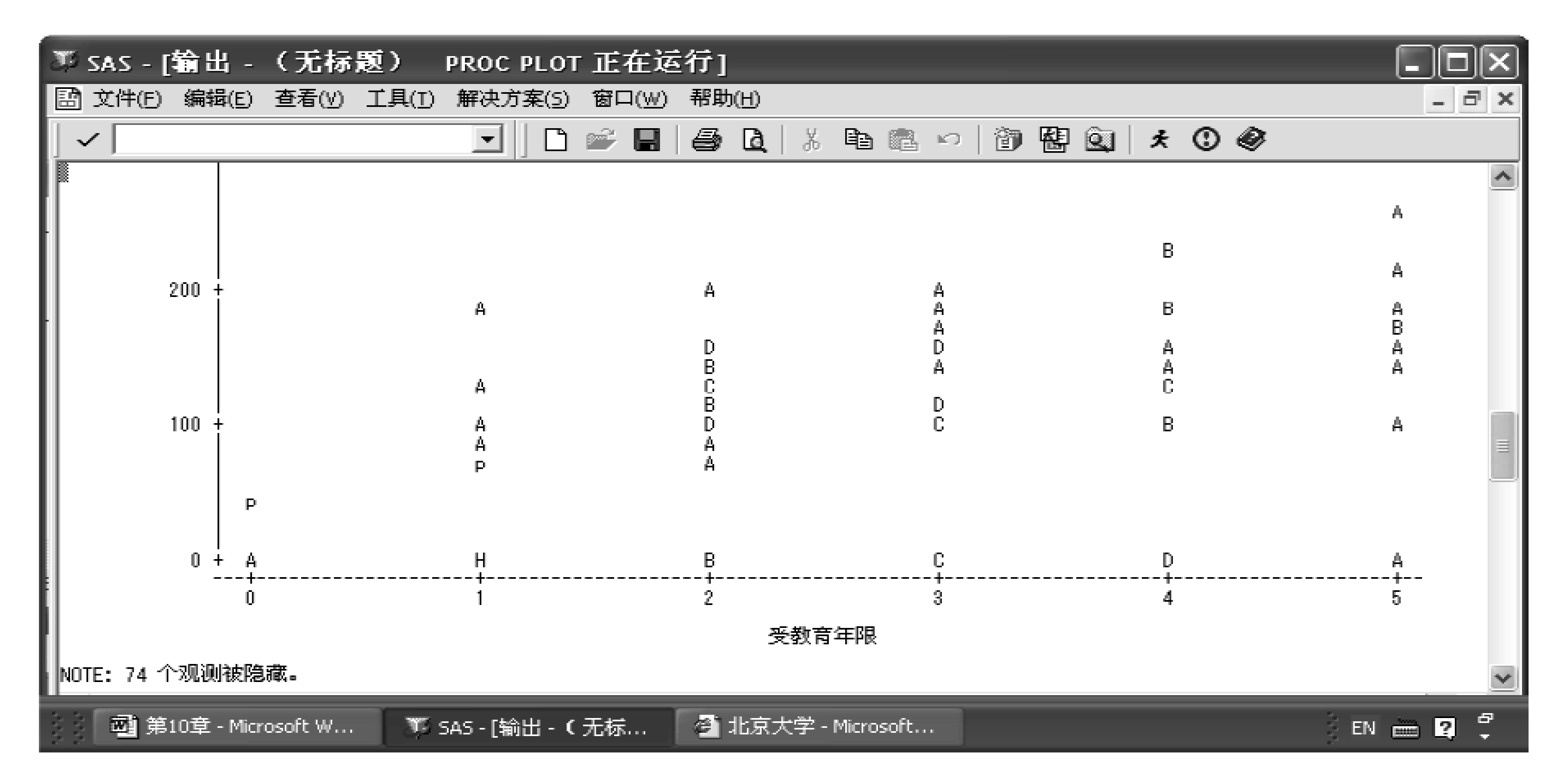


图 10.4 PLOT 工资 * 教育水平的散点图

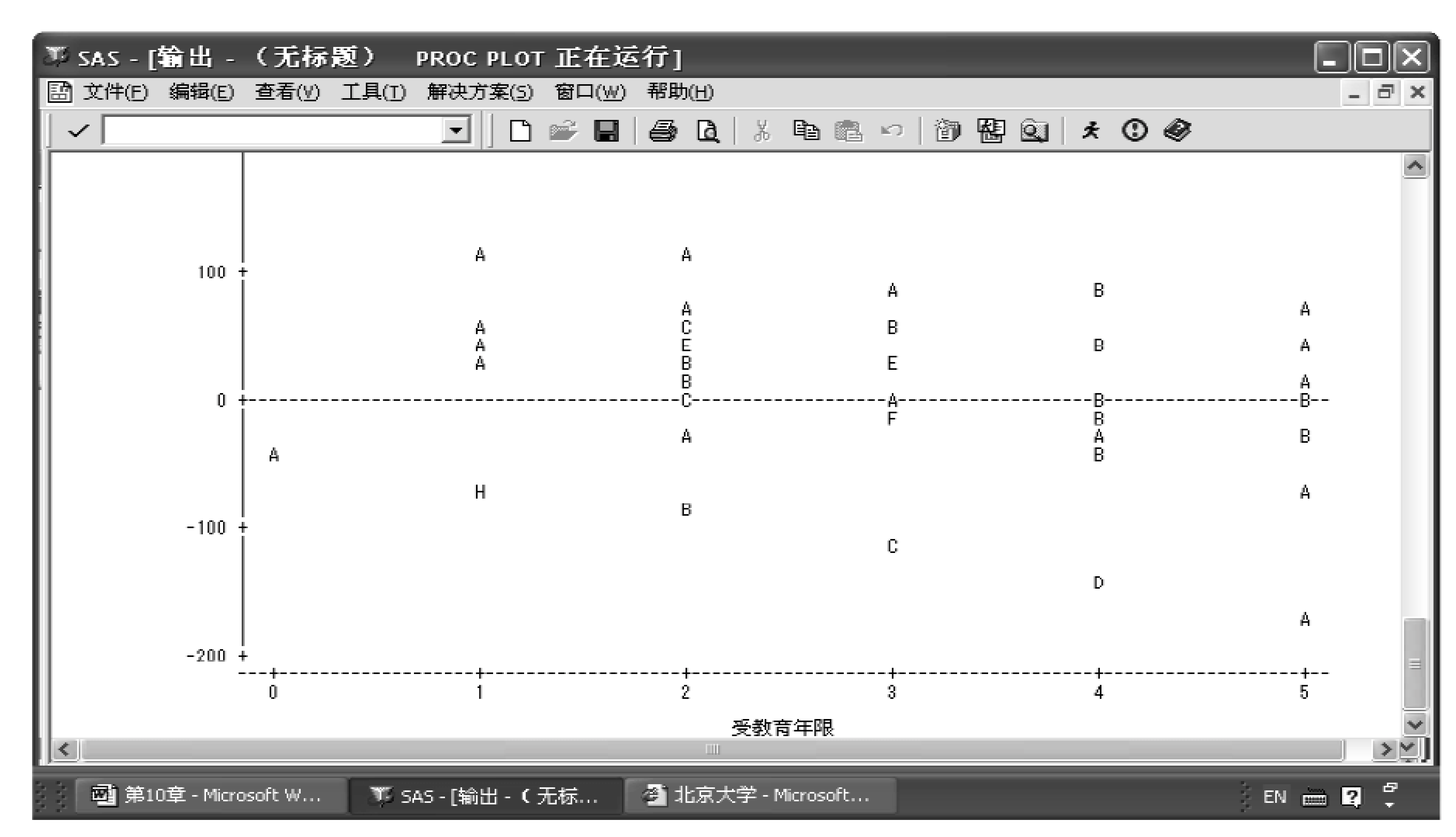


图 10.5 PLOT 残差 * 教育水平的散点图

10.4.2 数据统计

1. 北京市东城区社区服务研究数据分析之一：结果见图 10.3。
2. 大中城市社区服务研究《居民调查》数据分析之二：结果见图 10.4。
3. 大中城市社区服务研究《居民调查》数据分析之三：结果见图 10.5。

10.4.3 数据挖掘

下面逐一分析程序 10.1 的输出结果(见图 10.3～图 10.5)。

1. 程序 10.1 中的第三个 PROC 子程序

```
PROC  PLOT;
  PLOT  SAL2 * edc  PREDICT * EDC= 'P'/OVERLAY;
  PLOT  RESIDUAL * edc/VREF= 0;
```

这个“PROC PLOT;”子程序下面的第一个 PLOT 语句“PLOT SAL2 * edc PREDICT * EDC='P'/OVERLAY;”产生图 10.4。这是一幅“当前工资 * 教育水平(即：SAL2 * edc)”的散点图。等号“=”后面的 P 是任选字符，用于表示图点。再后面的“/OVERLAY”要求将两个图形叠在一起以便比较。

图 10.4 中的 A 代表 1 个观察值(人)，B 代表 2 个观察值(人)，S 代表 19 个观察值(OBS 即人，个案)等，余者类推。横轴是受教育年限(从文盲到大专以上)，纵轴是当前工资(由 0 到 1000 元)。从散点图 10.4 可知，小学文化程度的月收入 200 元左右的有 1 人(见图点 A)，初中文化程度的月收入 200 元的有 1 人(见图点“A”)；余者类推。

“PROC PLOT;”子程序的第二个“PLOT RESIDUAL * edc/VREF=0;”语句，产生图 10.5。这是一幅“残差 * 教育水平”的散点图。图中通过 0 点的一条直线是由该语句后面的“/VREF=0”产生的，表示“正负值的分界线”。但是，SAS 8.1 版本画出的“通过 0 点的一条直线”不是直线，而是由汉字组成的线，图 10.5 是 SAS 9e 版本产生的散点图。

通过散点图看出：残差是随着教育水平的提高而增加，说明残差不是独立的，模型不拟合数据。

2. 程序 10.1 中的“PROC GLM”子程序

运行 PROC GLM 整个子程序产生图 10.3 的结果。下面从 10 个方面加以分析。

(1) 方差分析的统计量。

Model 的 Sum of Squares：回归平方和(Regression Sum of Squares 简称 SSR)=90757.7391。误差平方和(Error Sum of Squares 简称 SSE)=654868.1951。总平方和(Total Sum of Squares 简称 SST)=745625.9342。三者的自由度分别为 1、74、75。回归平方和的均值(MSR)= 90757.7391，误差平方和的均值(MSE)=8849.5702。实际的 F 值=10.26，其概率 P 为 0.0020，小于显著性水平 $\alpha=0.05$。说明模型拟合数据，不必修改回归模型。

(2) 判定系数 R-Square：R-Square 即 R^2，其值为 0.121720。

因为 R^2 = SSR ÷ SST = 90757.7391 ÷ 745625.9342 = 0.121720，所以说明当前月收入的变化仅有 12.2%是来自教育水平的变化。因此，本例中用教育水平来预测退休后的月收入是不适合的，应选用其他自变量。

(3) Root MSE：等于 SQRT(MSE) = SQRT(8849.5702) = 94.07215。

(4) C·V 值：偏差(Coefficient of Variation)，表示总体的变异性，是由下式得到的：

$$C \cdot V = (\text{ROOT MSE}/\text{MEAN}) * 100$$

$$= (8849.5702/113.6184211) * 100 = 82.79657$$

(5) TYPE Ⅰ SS：表示第一类型的平方和，即：每一个自变量分别入选到回归模型时 MODEL 的平方和(SSR)的增值。其算法是先计算 MODEL 中的第一个自变量所解释的平方和，再计算第二个自变量所解释的剩余平方和。依此逐次计算出各个自变量的平方和—序列平方和。此例只有一个 edc 自变量，所以 TYPE Ⅰ SS(edc)等于原来的回归平方和 90757.7391。

(6) TYPE Ⅲ SS：表示第三类型的平方和，即当每一个自变量分别是最后一个入选到回归模型时，计算出的 MODEL 平方和(SSR)的增量。

此例只有一个 edc 自变量，所以 TYPE Ⅲ SS(edc)等于原来的回归平方和 90757.7391。

(7) Estimate：回归方程式及回归系数的估计值。由于截距项的回归系数 B_0 = 37.92682927，自变量 edc 的回归系数为 26.7560975，因此回归方程式为：

$$Y_i = 37.92682927 + 26.7560975 * X_i \tag{10.8}$$

(8) t Valu：用以检验各个自变量是否为 0 的 T 值。

H_0(原假设)：$B_0 = 0$；$B_1 = 0$

H_1(备择假设)：$B_0 \neq 0$；$B_1 \neq 0$

(9) Pr>|t|：这是第(8)项 T 值所对应的概率。本例的 Pr>|t|一栏下面分别为：0.0020。它小于显著性水平 $\alpha = 0.05$，说明与第(8)项的结论相同，即自变量的回归系数不能为 0。

(10) Standard Error：这是回归系数估计的标准误差。从图 10.3(b)可知，截距项的标准误差为 25.98237817，自变量 edc 的标准误差为 8.35491481，较大。

3. 分析摘要

了解以上各个参数的来历和重要性之后，如果要判断模型的优劣，可看(1)方差分析的统计量，特别是 P 值：它位于 Pr>F 一栏，其值为 0.0020，小于显著性水平 α 值 0.05。说明模型拟合数据，不必修改回归模型。如果要观察某变量进入方程后对于模型的贡献多大，可以观察“(2)判定系数”R^2，R^2 应该大于 0.5。

最后是通过“(9)Pr>|t|”的值，观察某变量的回归系数是否可靠，此值应该小于 $\alpha = 0.05$(或 $\alpha = 0.001$)。

4. 怎样写回归方程(数据的挖掘)

综上所述，本例的模型拟合数据，而且本例的回归方程为：

退休后的工资＝37.92682927＋26.7560975＊受教育年限。

10.5 调用GLM程序进行多元线性回归分析

上一节介绍过简单的线性回归，它只用一个自变量来解释因变量。但在实际运用中，为了正确反映因果关系，往往要利用两个以上的自变量来解释因变量，这种方法称为多元回归（Multiple Regrssion）。

1. GLM的源程序

GLM的源程序是在程序10.1的基础上建立多元回归模型（有edc和ocu1两个自变量）的。

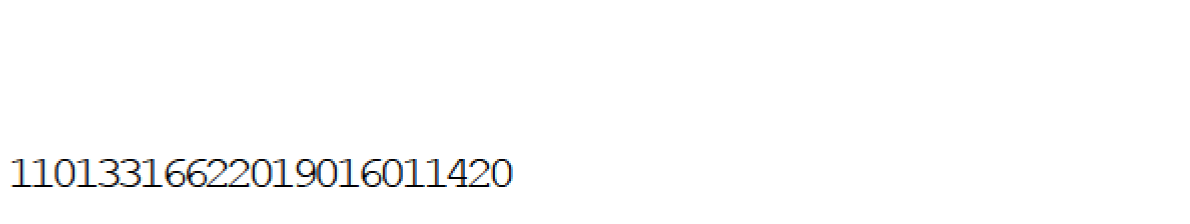
（1）二元回归语句

```
PROC GLM;
MODEL SAL2=edc ocu1;
```

（2）完整的程序10.2如下

程序10.2：

```
DATA Xsq1;
INPUT id1 1-2 caseid 3-5 n 6 sex 7 age 8-9 edc 10 ocu1 11 ocu2 12
      sal1 13-15 sal2 16-18 (v1-v5) (5*1.);
LABEL n='称谓' sex='老人性别' age='老人年龄' edc='受教育年限'
      ocu1='退休前职业' ocu2='退休后职业' sal1='退休前月收入:元'
      sal2='退休后月收入:元';
CARDS;
1100111774200781512311З
11001717942099172222215
1100252695200780991132l
11002617052015017011321
11003517452011018022222
11003626752011516812223
11004226121009007031311
11005525732620010011324
11006227521004610321310
11007228910000000031120
11008526921112515011210
11009617623007014012121
11010615842212010011314
11011526640000000031310
11012625142625014021414
11013117521609220011220
11013227411605010211320
```



```
1101331662201901 6011420
```

```
11042527946600000012222
11043517332615010011310
11044526216600000021410
11044616816600000021210
11045617742605013011425
11046728410000000032220
11047727010000000032212
11048525631020014011310
11048615531024000011210
11049116952208015011110
11050516922615013011313
11050626710000000032211
11051516842614023011210
11051626202000000031311
11082516235509010021210
12002616452619522011211
12003525821204006030010
12004517011608506013113
12004626121206405011411
12005617152636036013110
12006516132233033011210
12007116726600000011310
12007226016600000031310
12008428326600000031210
12009116122212531131121
12009225811606007831322
12010525621613012011411
12010615621613513011411
12011315826600000011220
12011425521600000011320
12012516221630017011311
12012626326600000021113
12013525846623028011210
12013615836623028011315
12014625821100000011315
12015616252630026011210
12016116636615000012224
12016225556615000031411
12017427816600000032110
12018117411618012021212
12018226811610008021314
12019526711603712511210
12020516746020020011310
12020626326015020011310
```

```
1202111605660000001132
12021225846600000021325
12022616552612000011310
12022526332609900011310
12023117311600005013110
12023226611600005011410
12024616355600020011212
12025316542600000011410
12025426042600000011410
12026116126620025011322
12026226321608010011324
12027316026620018012110
12027425436612008011310
12028516132233033011210
12029616455626025011312
12030116221018018011210
12030225421010509621310
12031225821118015011310
12031115842220000011212
12032527532606014012213
12033117326600017011320
12033227016600000032220
12034516936600021511210
12034617056600005032413
12035117321610000000000
12036226555620015011310
12037616242625025011310
12037525842618613811310
12038229216600000031210
12038626716600000031310
12039516352630030012113
12040618052618030012110
12040527754608016021410
12041526316600000021312
12042228200000000021210
12043227500000001321222
12044517333605113811221
12045116921110008011210
12045226811100000021410
12046229000000000032110
12047517721609728312204
;
PROC GLM;
PROC FORMAT;
```

```
VALUE n 1="父亲" 2="母亲" 3="岳父(或公公)" 4="岳母(或婆婆)"
        5="户主" 6="爱人" 7="其他(爷爷奶奶等)";
VALUE sex 1="男" 2='女';
VALUE edc 1='文盲' 2='小学' 3='初中'
          4='高中或中专' 5='大专以上';
PROC GLM;
  MODEL SAL2=edc;
  OUTPUT out=out1 P=predict R=residual;
PROC PLOT;
  PLOT sal2 * edc predict * edc="p" /OVERLAY;
  PLOT residual * edc/vref=0;
RUN;
```

然后，在图 10.6 的程序编辑器中编辑修改程序 10.2。



图 10.6 完整且可以执行的二元回归程序

运行图 10.6 中的程序 10.2 产生图 10.7 所示的结果。

从程序 10.2 的 MODEL 语句，可以挖掘出以下的二元回归模型：

$$SAL_2 = B_0 + B_1 * edc + B_2 * ocu_1 \tag{10.9}$$

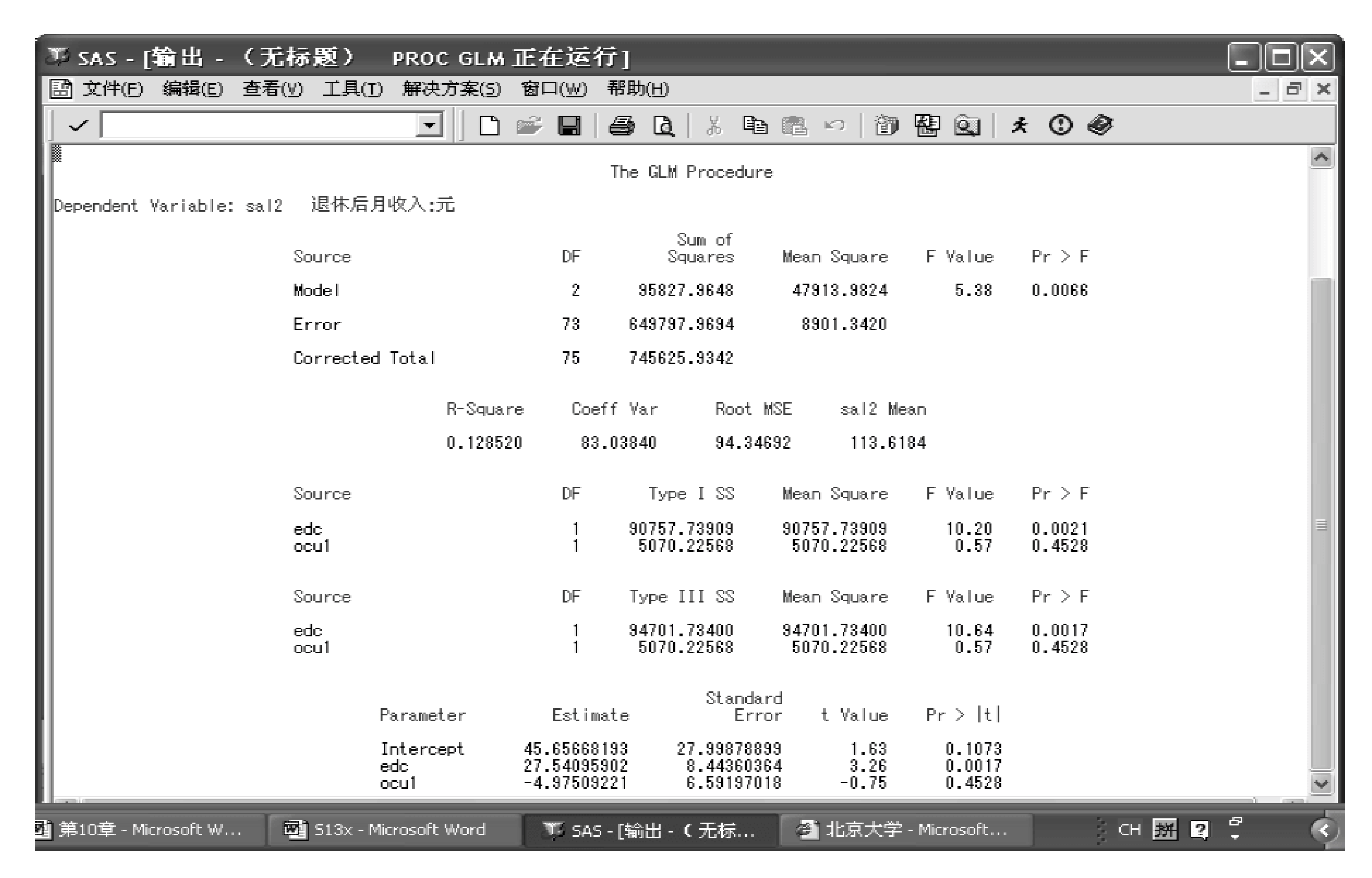
2. GLM 的输出结果

输出见图 10.7。

3. GLM 的结果分析

图 10.7 与图 10.3 大致相同。现把不同之处说明如下。

(1) Model 的 Sum of Squares：回归平方和为 95827.9648，因有两个自变量，因此自由度为 2。又因为观察值为 76 个，因此修正后的总平方和的自由度为(76－1)＝75，而误差平方和(Error sum of Squares)的自由度为 73。“Pr＞F”值为 0.0066 小于 α 值 0.05，



The GLM Procedure

Dependent Variable: sal2 退休后月收入:元

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	95827.9648	47913.9824	5.38	0.0066
Error	73	649797.9694	8901.3420		
Corrected Total	75	745625.9342			

R-Square	Coeff Var	Root MSE	sal2 Mean
0.128520	83.03840	94.34692	113.6184

Source	DF	Type I SS	Mean Square	F Value	Pr > F
edc	1	90757.73909	90757.73909	10.20	0.0021
ocu1	1	5070.22568	5070.22568	0.57	0.4528

Source	DF	Type III SS	Mean Square	F Value	Pr > F
edc	1	94701.73400	94701.73400	10.64	0.0017
ocu1	1	5070.22568	5070.22568	0.57	0.4528

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	45.65668193	27.99878899	1.63	0.1073
edc	27.54095902	8.44360364	3.26	0.0017
ocu1	-4.97509221	6.59197018	-0.75	0.4528

图 10.7 二元线性回归输出示意图

模型很显著。

(2) TYPE Ⅰ SS:是第一类型中第一个自变量 edc 所能解释的平方和,其值为 90757.73909(见 EDC 一行)。第二个自变量 ocu1 所解释的剩余平方和为 5070.22568。

可见,预先进入回归模型的自变量,比以后进入回归模型的自变量的平方和大得多。

(3) TYPE Ⅲ SS:表示,在计算每个自变量的平方和时,先排除其他自变量的影响。由此获得的平方和称为“偏平方和”(Partial sum of Squares),也称为第三类型的平方和。在本例中,eac 和 ocu1 自变量的偏平方和分别为:94701.73400 与 5070.22568。

(4) Pr>|t| :此项是“回归系数 $B_0=0$”的 t 值的检验值,变量 edc 的 t 检验的概率 P 值等于 0.0017,小于 α 值 0.05,变量 edc 的回归系数有效。但变量 ocu1 的 t 检验的概率 P 值等于 0.4528,不显著。

(5) Estimate:回归系数估计值。截距项的回归系数为 $B_0\approx45.66$,自变量 edc 的回归系数为 $B_1\approx27.54$。第二个自变量 ocu1 的回归系数为 $B_2\approx-4.98$,但因不显著而忽略。所以此例的回归方程式为:

$$Y_i=45.65668193+27.54095902*X$$

或

$$SAL_i = 45.66 + 27.54 * edc \qquad (10.10)$$

10.6 调用 GLM 程序进行多项式回归

10.6.1 多项式回归的一般模型

多项式回归与多元回归的相似之处,是由一个或两个以上自变量解释一个因变量。

其不同之处在于，多项式回归模型中的自变量可以以幂的方式表示。例如：一个自变量的二次方程为：

$$Y_i = B_0 + B_1 X_i + B_2 X_i^2$$

一个自变量的三次方程为：

$$Y_i = B_0 + B_1 X_i + B_2 X_i^2 + B_3 X_i^3$$

一个自变量的高次方程为：

$$Y_i = B_0 + B_1 X_i + B_2 X_i^2 + B_3 X_i^3 + \cdots + B_k X_i^k$$

同理：两个自变量的二次方程为：

$$Y_i = B_0 + B_1 X_{i1} + B_1 X_{i2} + B_{11} X_{i1}^2 + B_{12} X_{i1} X_{i2} + B_{22} X_{i2}^2$$

上述模型是一个或几个自变量解释一个因变量 Y。当自变量变化时，对 Y 的影响不仅在程度上，而且在方向上均产生变化。

10.6.2 多项式回归的实例

有一个厂家，5 个月中产品的广告费用，与销售额之间的抽样数据见表 10.4。

表 10.4 产品的广告费与销售额的数据

X：广告费用(元)	Y：销售额(元)	X：广告费用(元)	Y：销售额(元)
1,000	101,000	2,000	209,000
1,250	116,000	2,500	264,000
1,500	165,000		

1. 求解

(1) 销售额与广告费用，及销售额与(广告费用)^{2}的一元二次回归模型；

(2) 相关系数和判定系数；

(3) 是否有推论意义($\alpha = 0.05$)。

2. 解答

根据题意，建立以下的一元二次回归模型

$$Y = B_0 + B_1 X_1 + B_2 X_1 * X_1$$

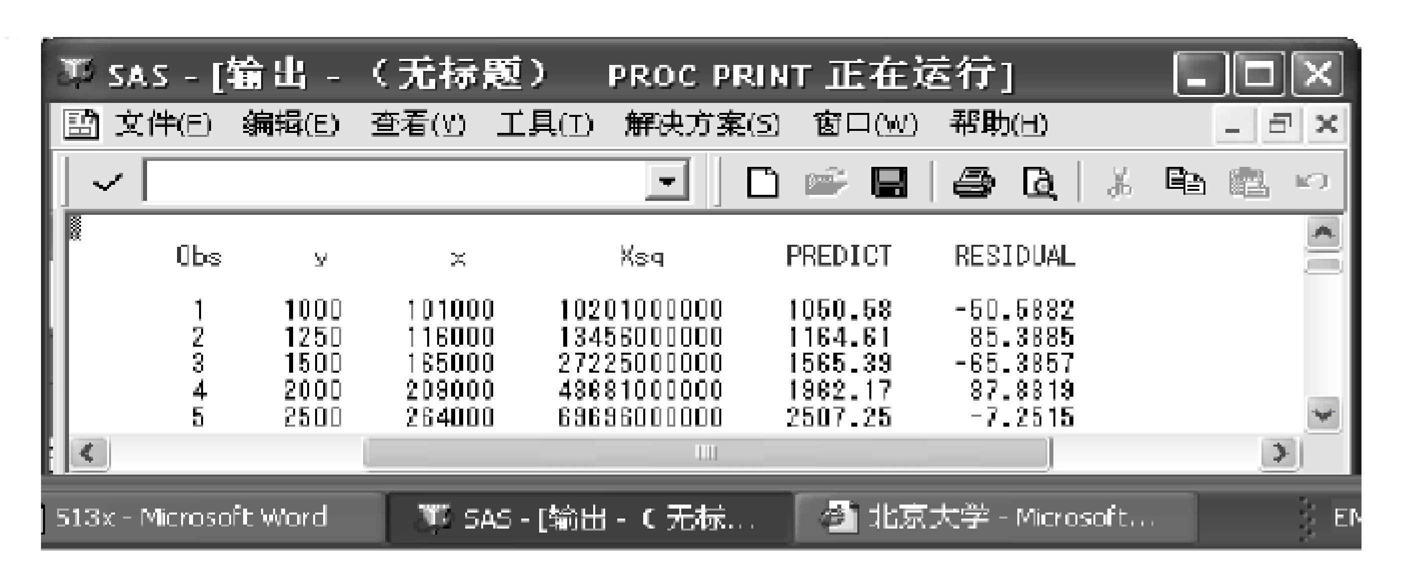
同时，设计了计算回归系数的多项式回归程序(见程序 10.3)。

程序 10.3：回归模型为 y＝aX＋bX＊X 的回归分析。

```
DATA sales;
INPUT y x @;
      Xsq=x* * 2;    /* 计算 X的平方值 * /
LIST;
CARD;
1000   101000
1250   116000
```

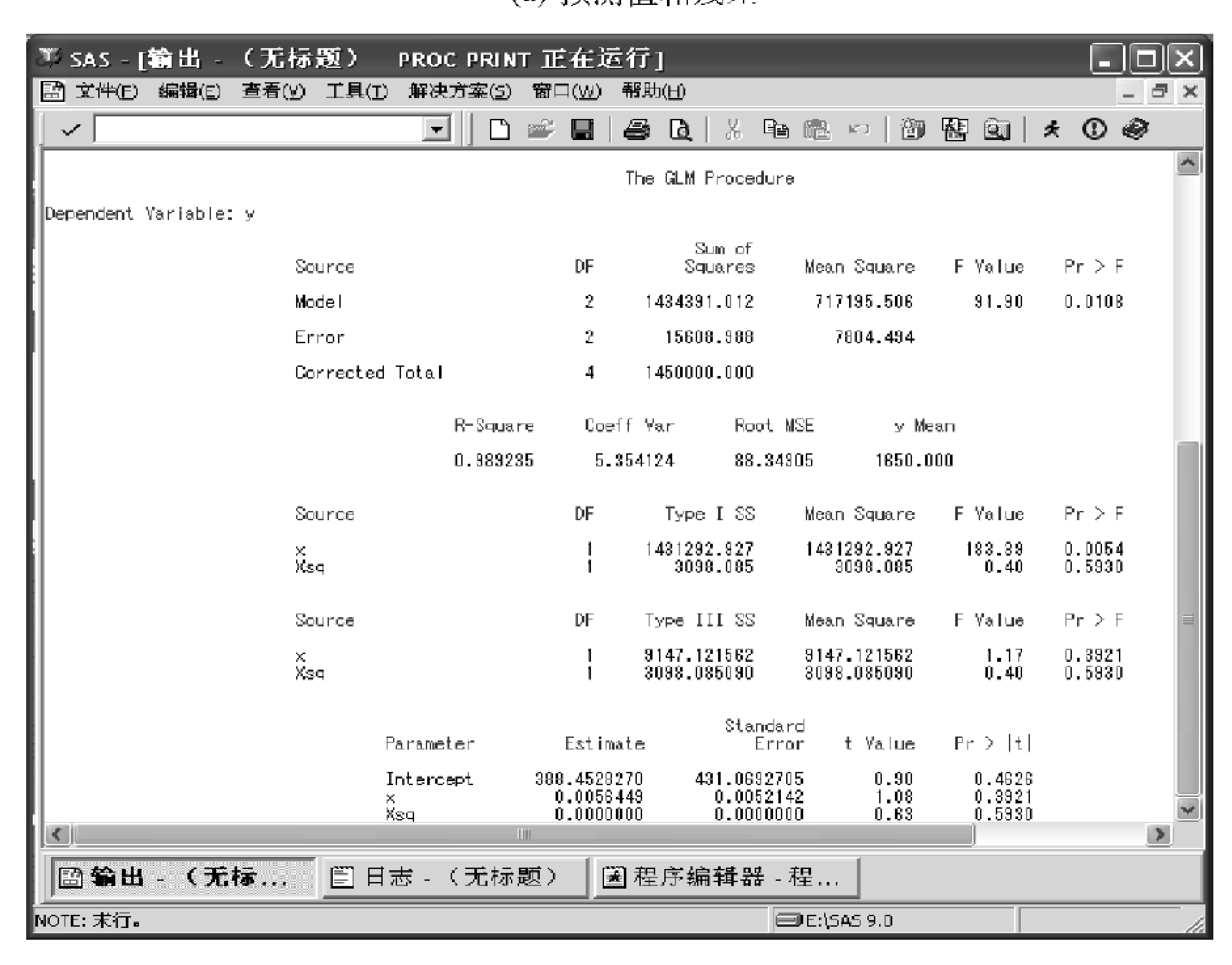
```
1500    165000
2000    209000
2500    264000
;
PROC print;                                 /* 显示 Y,X,X的平方值 * /
PROC GLM;
  MODEL y= x xsq;                           /* y为因变量,X和 Xsq为自变量 * /
  OUTPUT OUT= sa P= PREDICT R= RESIDUAL;    /* 可省 * /
PROC PRINT DATA= sa;
```

运行程序 10.3 产生图 10.8 所示的结果。



Obs	y	x	Xsq	PREDICT	RESIDUAL
1	1000	101000	10201000000	1050.58	-50.5882
2	1250	116000	13456000000	1164.61	85.3885
3	1500	165000	27225000000	1565.39	-65.3857
4	2000	209000	43681000000	1962.17	37.8319
5	2500	264000	69696000000	2507.25	-7.2515

(a) 预测值和残差



The GLM Procedure

Dependent Variable: y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1434391.012	717195.506	91.90	0.0108
Error	2	15608.988	7804.494		
Corrected Total	4	1450000.000			

R-Square	Coeff Var	Root MSE	y Mean
0.989235	5.354124	88.34305	1650.000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
x	1	1431292.927	1431292.927	183.39	0.0054
Xsq	1	3098.085	3098.085	0.40	0.5930

Source	DF	Type III SS	Mean Square	F Value	Pr > F
x	1	9147.121562	9147.121562	1.17	0.3921
Xsq	1	3098.085090	3098.085090	0.40	0.5930

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	388.4528270	431.0692705	0.90	0.4626
x	0.0056449	0.0052142	1.08	0.3921
Xsq	0.0000000	0.0000000	0.63	0.5930

(b) 回归模型

图 10.8　一元二次（$Y=ax+bx^2$）回归模型的输出示意图

3. 多项式回归的输出结果

图 10.8 是多项式回归的输出结果。

4. 结果分析

(1) H_0(原假设):$B_1=B_2=B_3=\cdots\cdots=B_i=0$(回归系数全为 0)。

从图 10.8(b)得知,由于“F Value”是模型检验项,此值的概率 P 为 0.0108,小于 α 值0.05,所以拒绝 H_0 原假设,回归系数不为 0,说明回归模型正确。

(2) Estimate:估算的回归系数。本例中,该截距项的回归系数虽为 $B_0=388.4528270$,但因不显著而被排除。再看自变量 X 与 XSQ 的回归系数,它们分别为 $B_1=0.0056449$ 和 $B_2=0.0000000$(也可以忽略不计),因而本例的回归方程式虽然为:销售额 $Y=388.4528270+0.0056449*$广告量,但是二次项完全不适合。

(3) t Value:T 分布,用于检验回归系数是否显著。例如截距 B_0(Intercept)的回归系数 t 值等于 0.9,此值小于 T(0.05,2)=2.92。或者说显著性水平 0.4626 大于 α(0.05),所以没有充分的理由拒绝原假设(注意:此时不能说成接受原假设),说明回归模型的截距项与 0 的差别不显著。本例没有推论意义。

(4) $R^2=0.989235$,接近百分百,说明自变量 X 解释了因变量 98%的变异性,判定系数很好。所以不需要设置 X 的平方项 Xsq。

综上所述本例数据只适合于一元一次回归模型。

10.7 虚拟变量的用法

通常情况下,回归分析中所用的自变量的数值是连续的,但有时有必要创建一个(或几个)虚拟变量,将原有的变量值表示成两个或几个特殊的数值。例如,考生的成绩,如果政治和数学的成绩都达到 85 分者,这个虚拟变量(Dummy Variable)的值为 1,否则为 0。

下面举个例子。考生语文、数学、英语等科目的平时成绩见数据文件 score.dat,用虚拟变量计算回归参数及其绘图的程序见程序 10.4。

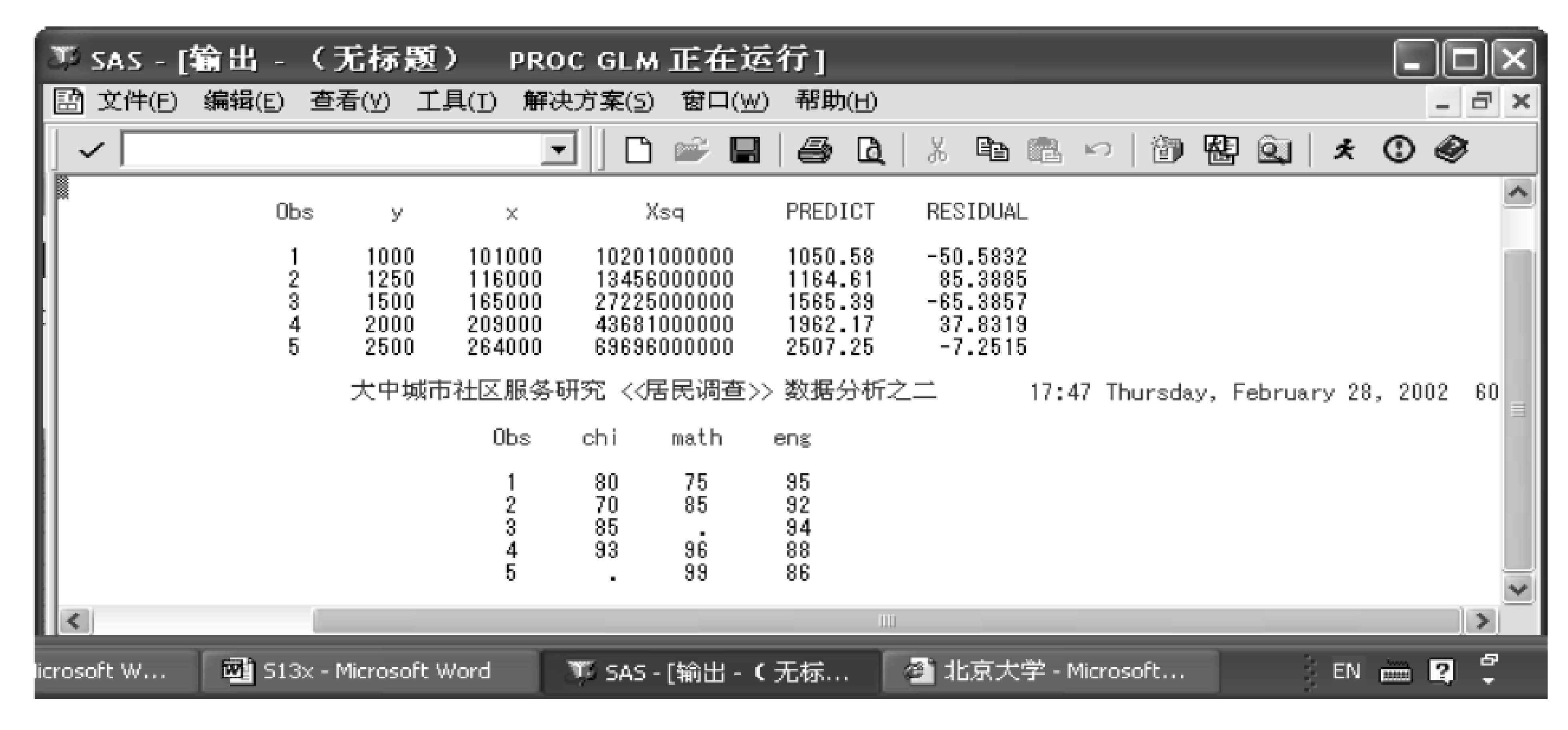
1. 虚拟变量的简例

程序 10.4:虚拟变量。

```
DATA score;
INPUT chi math eng;
LABLE chi='语文成绩' math='数学成绩' eng='英语成绩' schi='全班语文总分';
cards;
80   75   95
70   85   92
85   .    94
```

```
93   96   88
.    99   86
;
If chi>80|math>85 then dummy=1;
   else dummy=0;
list;                    /* 列出数据 */
PROC PRINT;              /* 显示统计结果 */
Proc Glm;
  model eng=chi dummy;
RUN;
```

运行程序10.4产生图10.9所示的结果。



Obs	y	x	Xsq	PREDICT	RESIDUAL
1	1000	101000	10201000000	1050.58	-50.5832
2	1250	116000	13456000000	1164.61	85.3885
3	1500	165000	27225000000	1565.39	-65.3857
4	2000	209000	43681000000	1962.17	37.8319
5	2500	264000	69696000000	2507.25	-7.2515

Obs	chi	math	eng
1	80	75	95
2	70	85	92
3	85	.	94
4	93	96	88
5	.	99	86

图10.9　虚拟变量的分析

2. 虚拟变量的输出结果

看程序10.4中IF语句的另一种等效表示法：

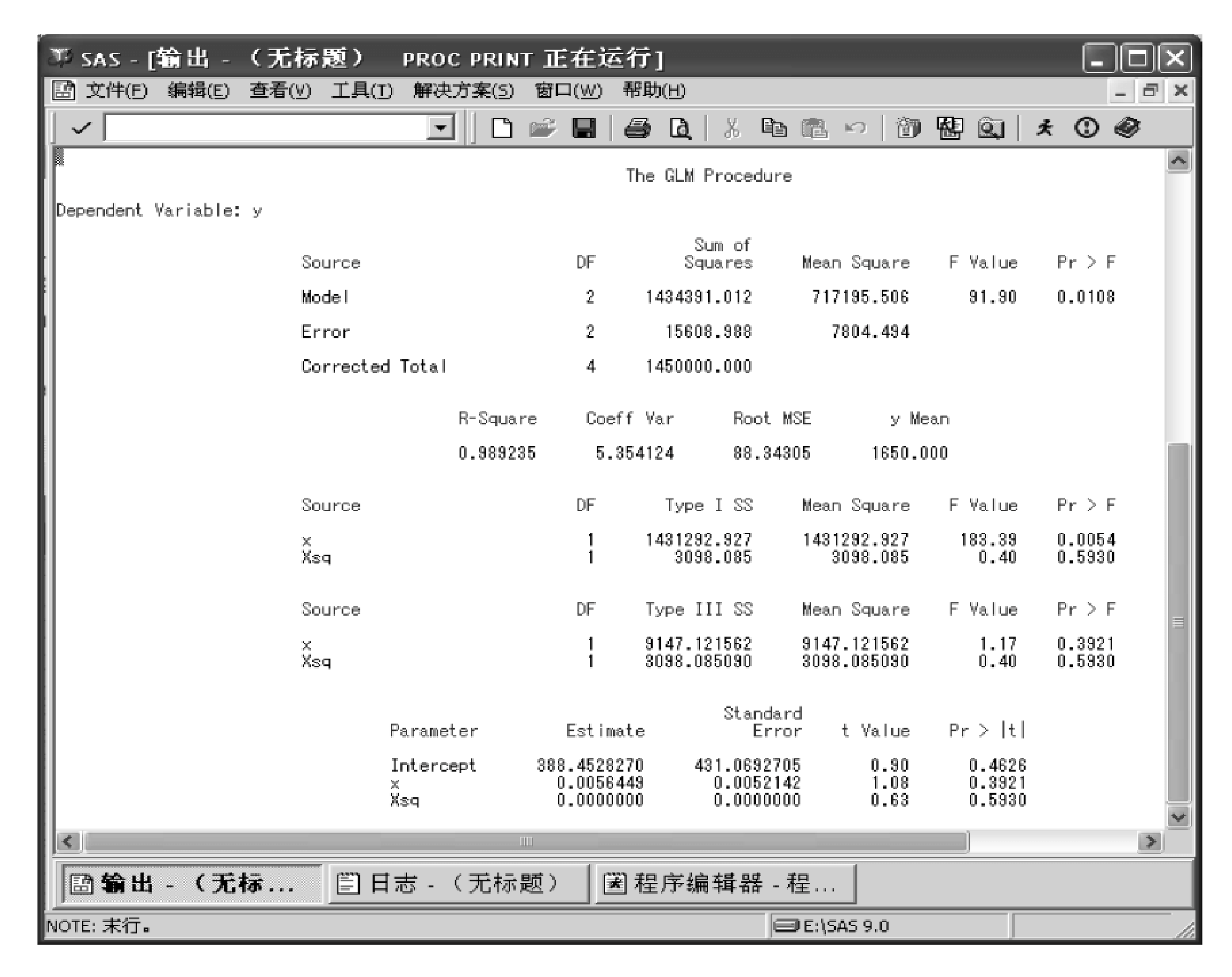
```
IF chi GT 85 OR Math GT 75 THEN Dummy=1;
ELSE Dummy=0;
```

这是设置虚拟变量DUMMY，表示当语文成绩大于85分，或者数学成绩大于75分时，虚拟变量Dummy值为1，否则为0。

3. 关于图10.9中各参数的分析，请参考图10.8的结果分析。

习　题　10

1. GLM过程有哪些功能？
2. 从图10.10和图10.11看，有无必要创建 $X*X$ 项？



SAS - [输出 - (无标题) PROC PRINT 正在运行]

The GLM Procedure

Dependent Variable: y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1434391.012	717195.506	91.90	0.0108
Error	2	15608.988	7804.494		
Corrected Total	4	1450000.000			

R-Square	Coeff Var	Root MSE	y Mean
0.989235	5.354124	88.34305	1650.000

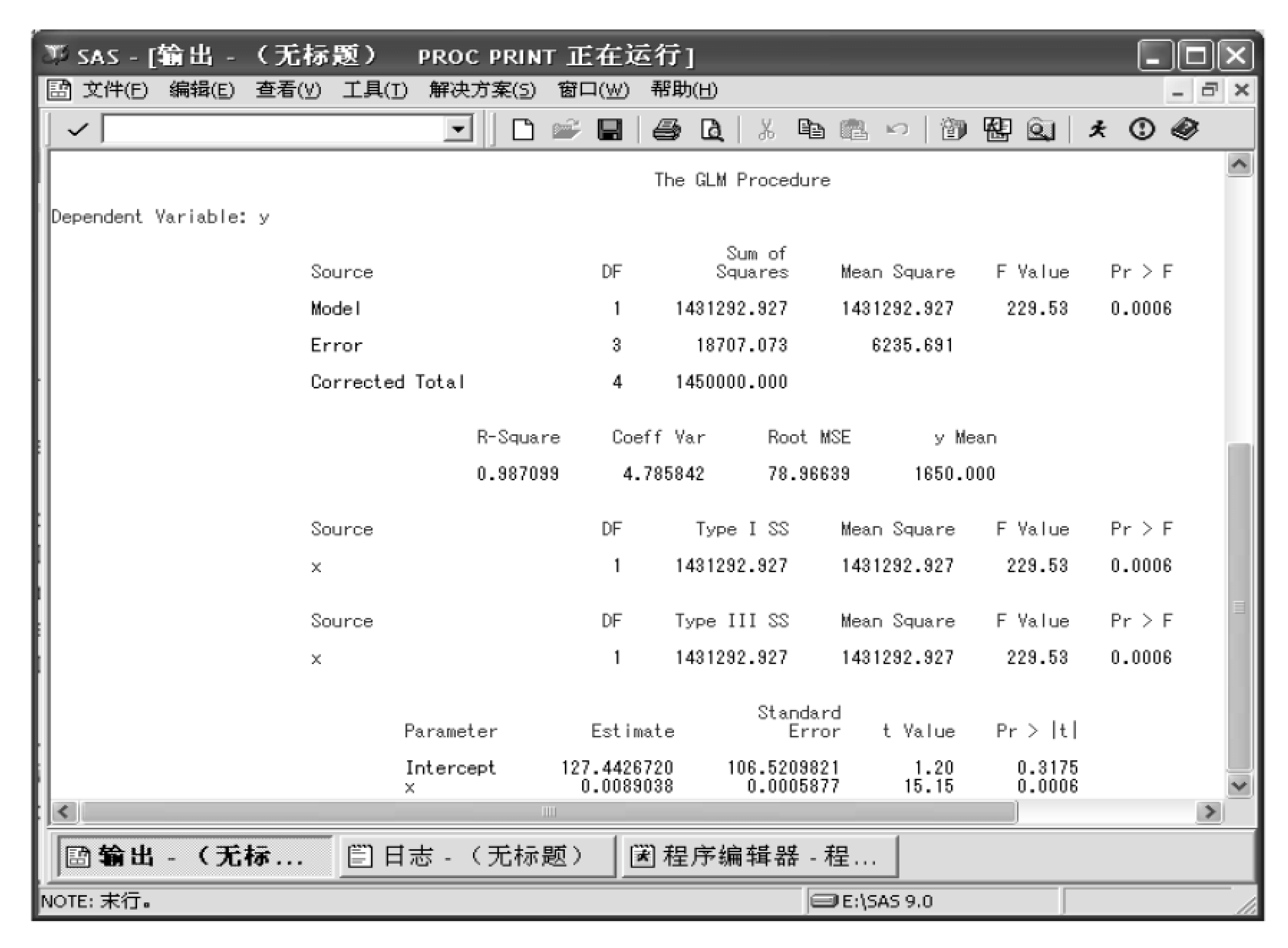
Source	DF	Type I SS	Mean Square	F Value	Pr > F
x	1	1431292.927	1431292.927	183.39	0.0054
Xsq	1	3098.085	3098.085	0.40	0.5930

Source	DF	Type III SS	Mean Square	F Value	Pr > F
x	1	9147.121562	9147.121562	1.17	0.3921
Xsq	1	3098.085090	3098.085090	0.40	0.5930

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	388.4528270	431.0692705	0.90	0.4626
x	0.0056449	0.0052142	1.08	0.3921
Xsq	0.0000000	0.0000000	0.63	0.5930

NOTE: 末行。 E:\SAS 9.0

图 10.10 含有 $X*X$ 项的回归结果



SAS - [输出 - (无标题) PROC PRINT 正在运行]

The GLM Procedure

Dependent Variable: y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1431292.927	1431292.927	229.53	0.0006
Error	3	18707.073	6235.691		
Corrected Total	4	1450000.000			

R-Square	Coeff Var	Root MSE	y Mean
0.987099	4.785842	78.96639	1650.000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
x	1	1431292.927	1431292.927	229.53	0.0006

Source	DF	Type III SS	Mean Square	F Value	Pr > F
x	1	1431292.927	1431292.927	229.53	0.0006

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	127.4426720	106.5209821	1.20	0.3175
x	0.0089038	0.0005877	15.15	0.0006

NOTE: 末行。 E:\SAS 9.0

图 10.11 模型中删除 $X*X$ 项后的回归输出

第 11 章 chapter 11

采用 REG 过程进行多元线性回归分析

在大、中、小型机的 SAS 系统中，REG、GLM、RSQUARE、STEPWISE 等各个回归过程都是分别独立的。但微型计算机 SAS 系统的 REG 过程中则包含了 RSQUARE（判定法）、STEPWISE（逐步回归法）、Forward Selection（向前选择法）、Backward dimination（自后淘汰法）等回归法；而 GLM 过程是一个单独存在的回归过程。

本章介绍 REG 过程的回归应用。

通常由一组自变量可组建多个回归模型。例如，有 3 个自变量 X1、X2、X3 和一个因变量 Y，则可组建（$2^3-1=$）7 种回归模型，即：

Y 对应 X1，X2，X3

Y 对应[X1，X2][X1，X3][X2，X3]

Y 对应[X1，X2，X3]

又例如，有 10 个自变量 X1－X10 和一个因变量时，则可组建 $2^{10}-1=1023$ 个回归模型。

因此，遇到多个自变量时，就应考虑哪些变量放在回归模型中为宜。

在实际工作中，最常用的回归法有：逐步回归法、向前选择法、自后淘汰法以及判定法。究竟选择哪一种回归法，一般应根据抽样的数据类型来选择其中的一些自变量及因变量，进而预测出最佳的回归模型。

而且，在逐步回归分析的程序中，一般应利用两种回归命令，一种是 RSQUARE，另一种是 STEPWISE。

下面以程序 11.1 中的数据及数据定义语句为例，分别采用对话框（见 11.1 节）和命令（见 11.3 节）进行回归分析。

程序 11.1：《社区服务研究》数据及数据定义语句，见图 11.1。

本例的数据有 474 个观察值（人），因变量为 vi（家庭月收入）；自变量为 v7f（全家总人口）和 wk（具体工作）等。



图 11.1 程序 11.1 中的《社区服务研究》数据及数据定义语句

11.1 用 Analyst 对话框做多元线性回归

1. Analyst(分析家)的操作步骤

(1) 选择图 11.1(第 1 行的 SAS 主菜单)中的"运行"→"提交"(或 Run→Submit)命令,运行程序 11.1 产生 SAS 数据集 sq14。

(2) 选择 SAS 主菜单中的"解决方案"→"分析"→"分析家"(Solutions→Analysis→Analyst)命令,如图 11.2 所示。



(a) Analyst的菜单位置

图 11.2 打开 SAS 工作区的文件 Work. sq14

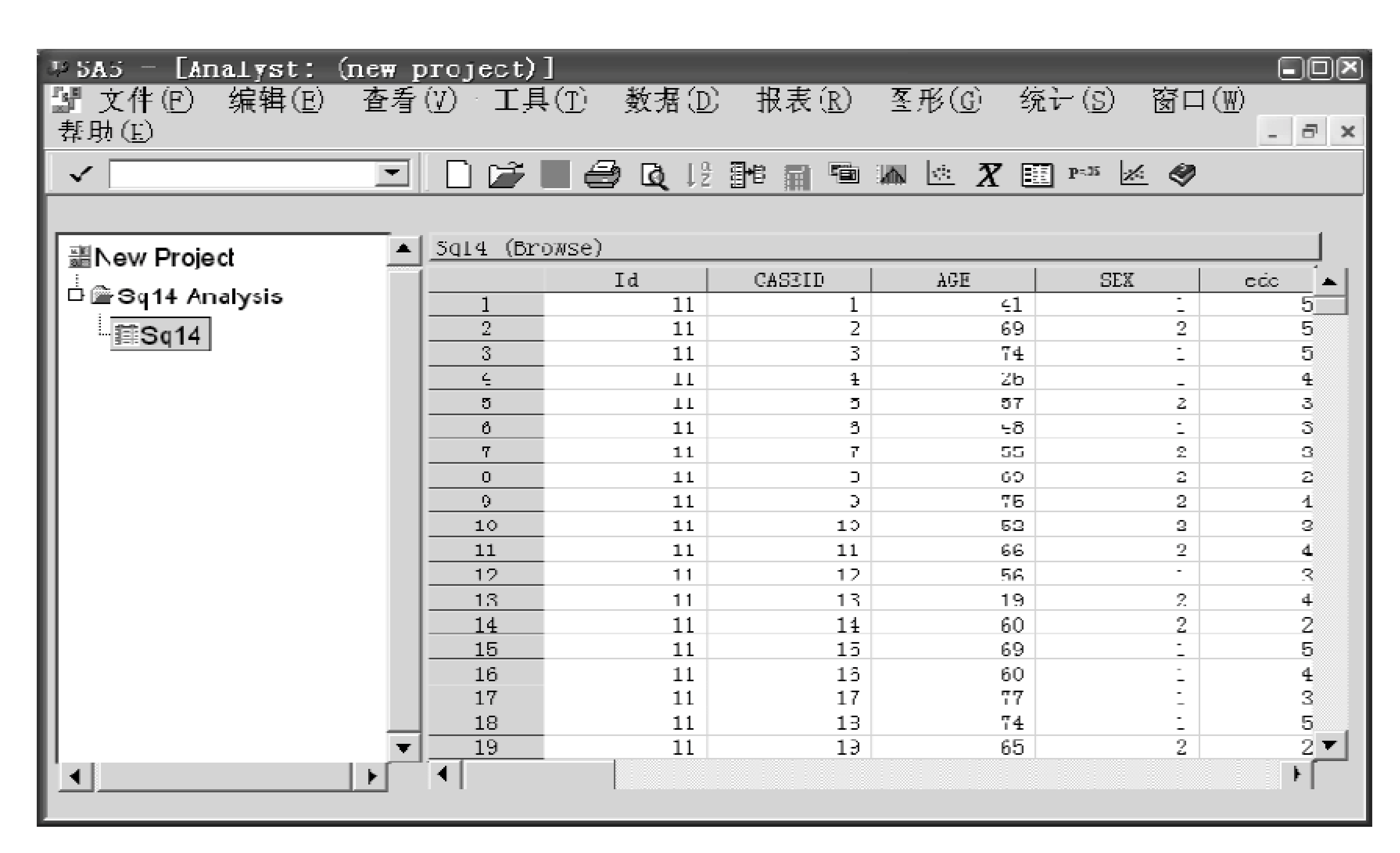


(b) Work.sq14 工作文件

图 11.2（续）

（3）选择“分析家”→“文件”→“按 SAS 文件名称打开”→Work（或 Analyst→File→Open By SAS Name→Work）命令，进入图 11.2(b)。

（4）选择文件名 sq14 后单击“确定”按钮，SAS 自动展示 Work. sq14 数据集的内容，见图 11.3。



	Id	CASEID	AGE	SEX	edc
1	11	1	41	1	5
2	11	2	69	2	5
3	11	3	74	1	5
4	11	4	26	1	4
5	11	5	57	2	3
6	11	6	48	1	3
7	11	7	55	2	3
8	11	8	60	2	2
9	11	9	76	2	1
10	11	10	53	2	3
11	11	11	66	2	4
12	11	12	56	1	3
13	11	13	19	2	4
14	11	14	60	2	2
15	11	15	69	1	5
16	11	16	60	1	4
17	11	17	77	1	3
18	11	18	74	1	5
19	11	19	65	2	2

图 11.3　Work. sq14 文件的内容（部分）

（5）选择“统计”→“回归”（或 Statistics→Regression）命令，展示图 11.4。

（6）选择“线性”（或 Linear）命令，进入图 11.5 并设置变量。

（7）单击 Model，进入图 11.6 并选择逐步回归法。

（8）单击图 11.6 中的 Statistics 标签进入图 11.7，并选择 Adjusted R-square 法。

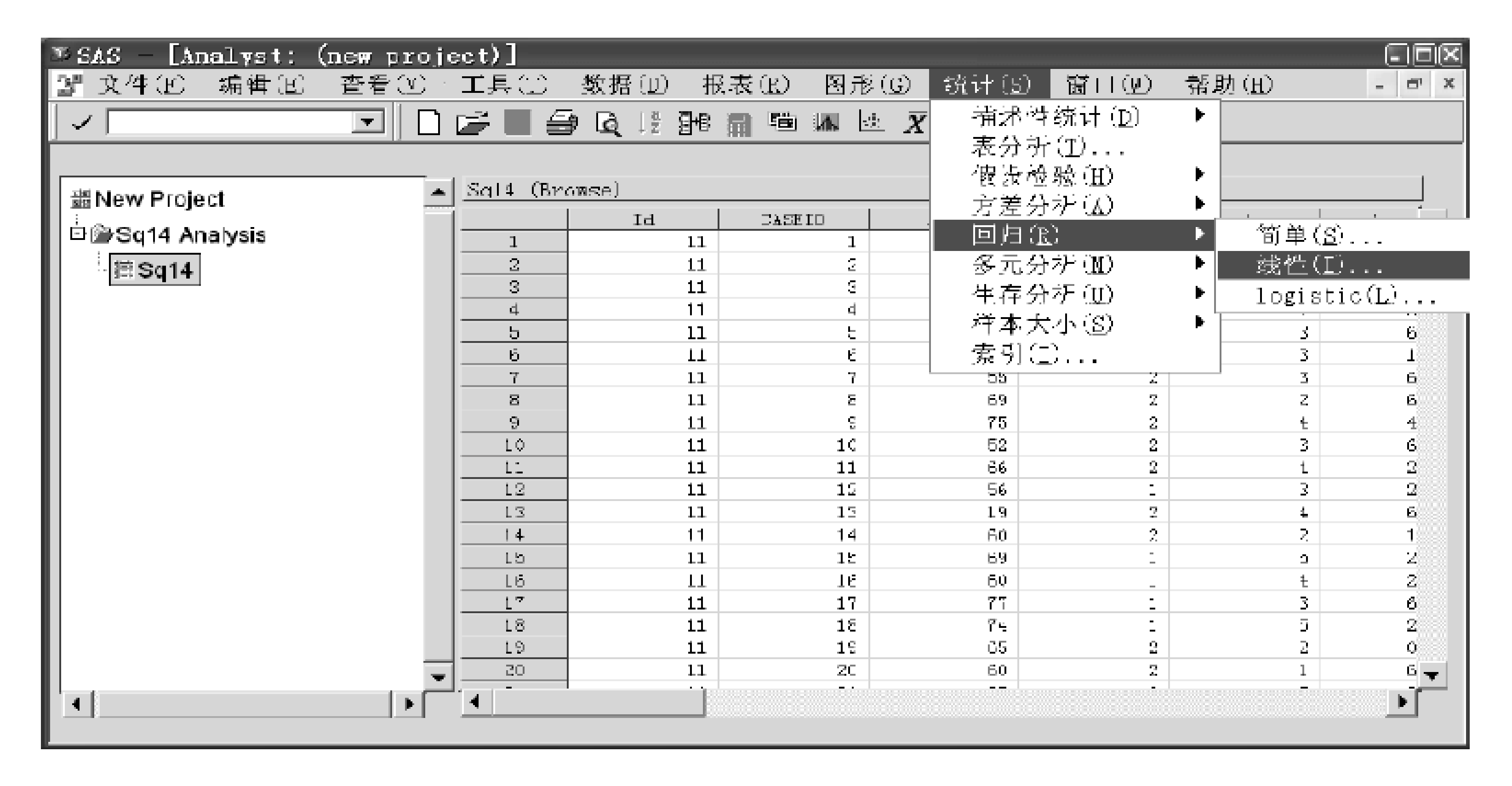


图 11.4　Linear Regression 的菜单位置

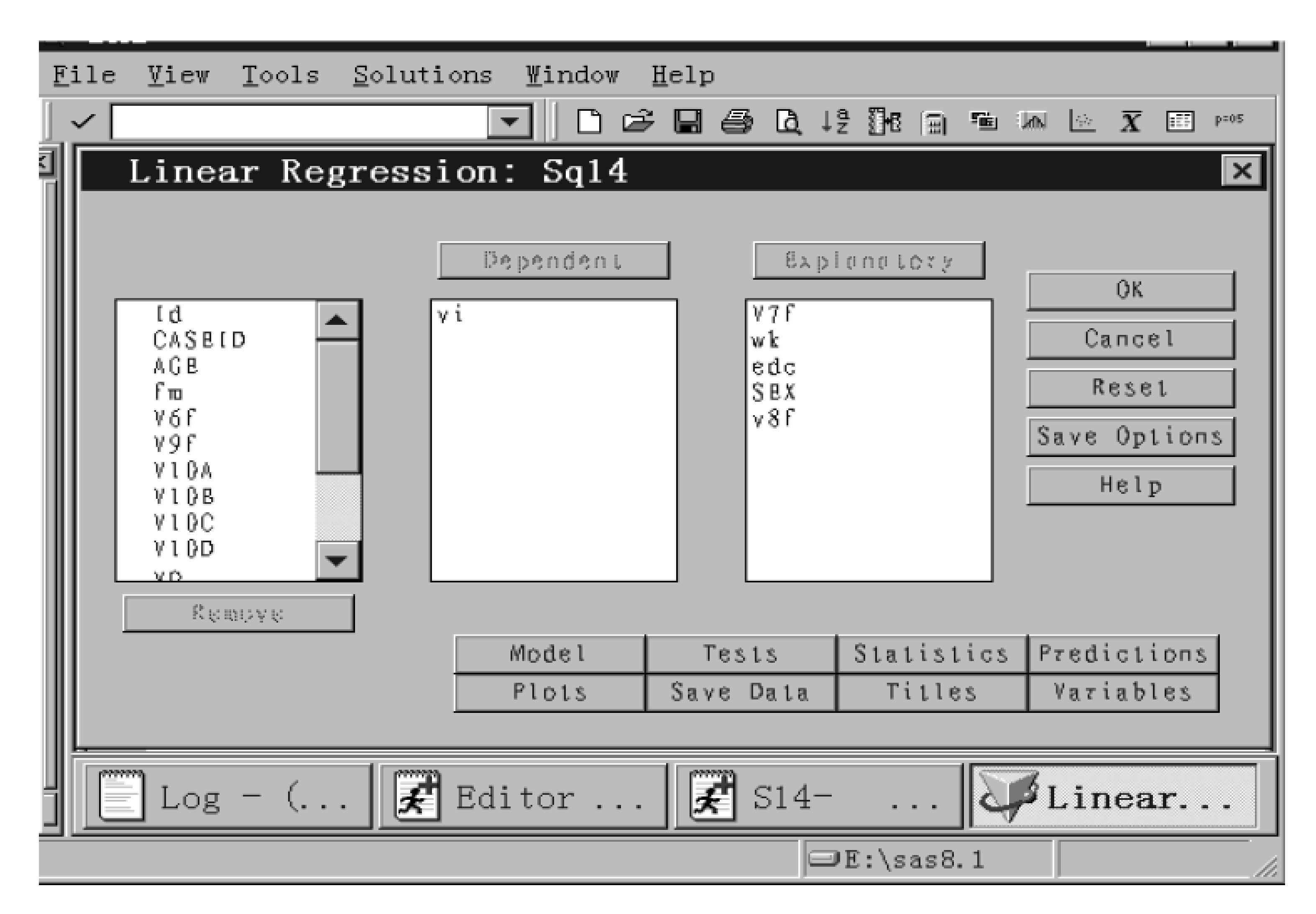


图 11.5　选择回归模型中的因变量 vi(月收入)和自变量 v7f、wk 等

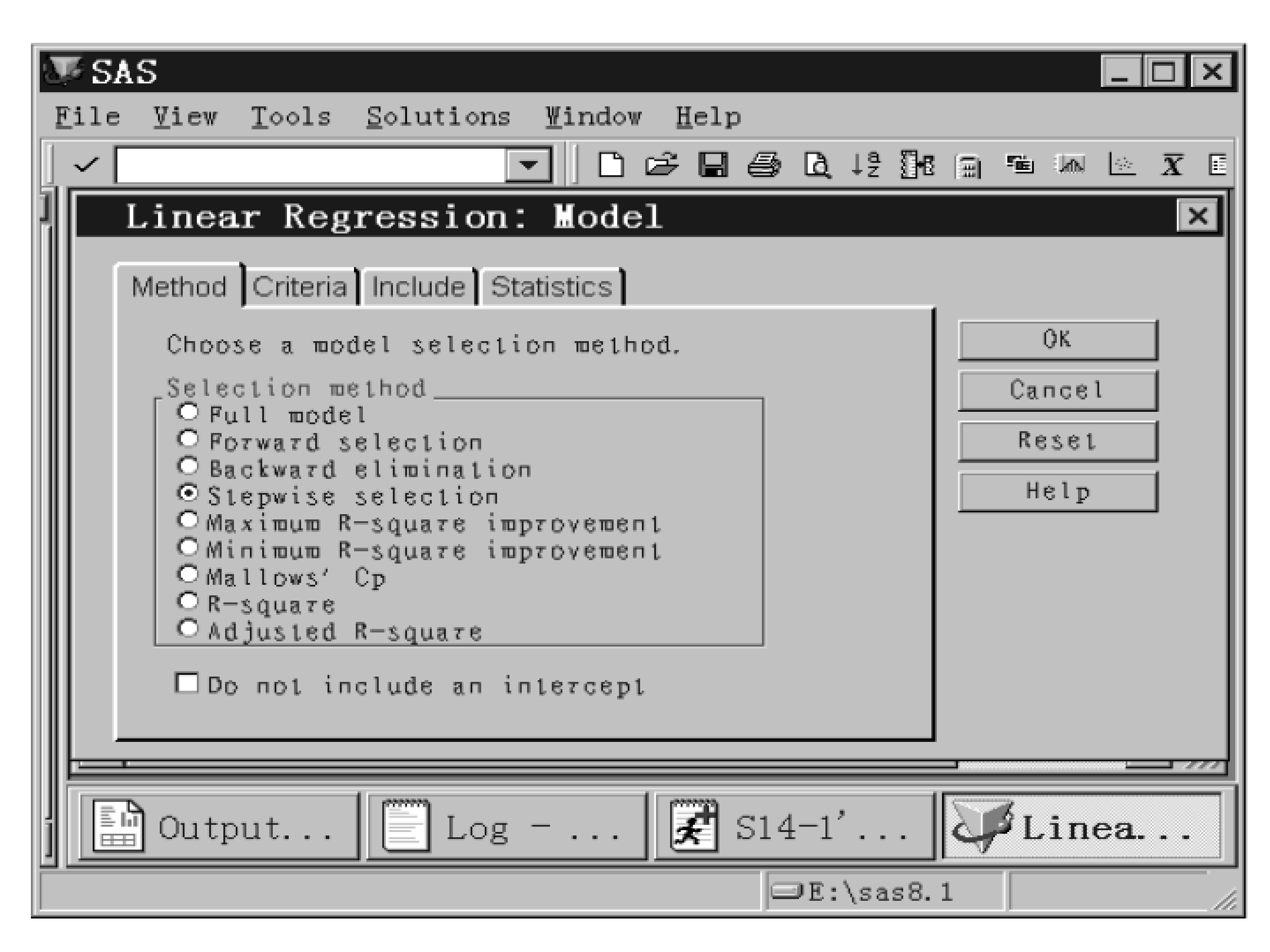


图 11.6　选择逐步回归法

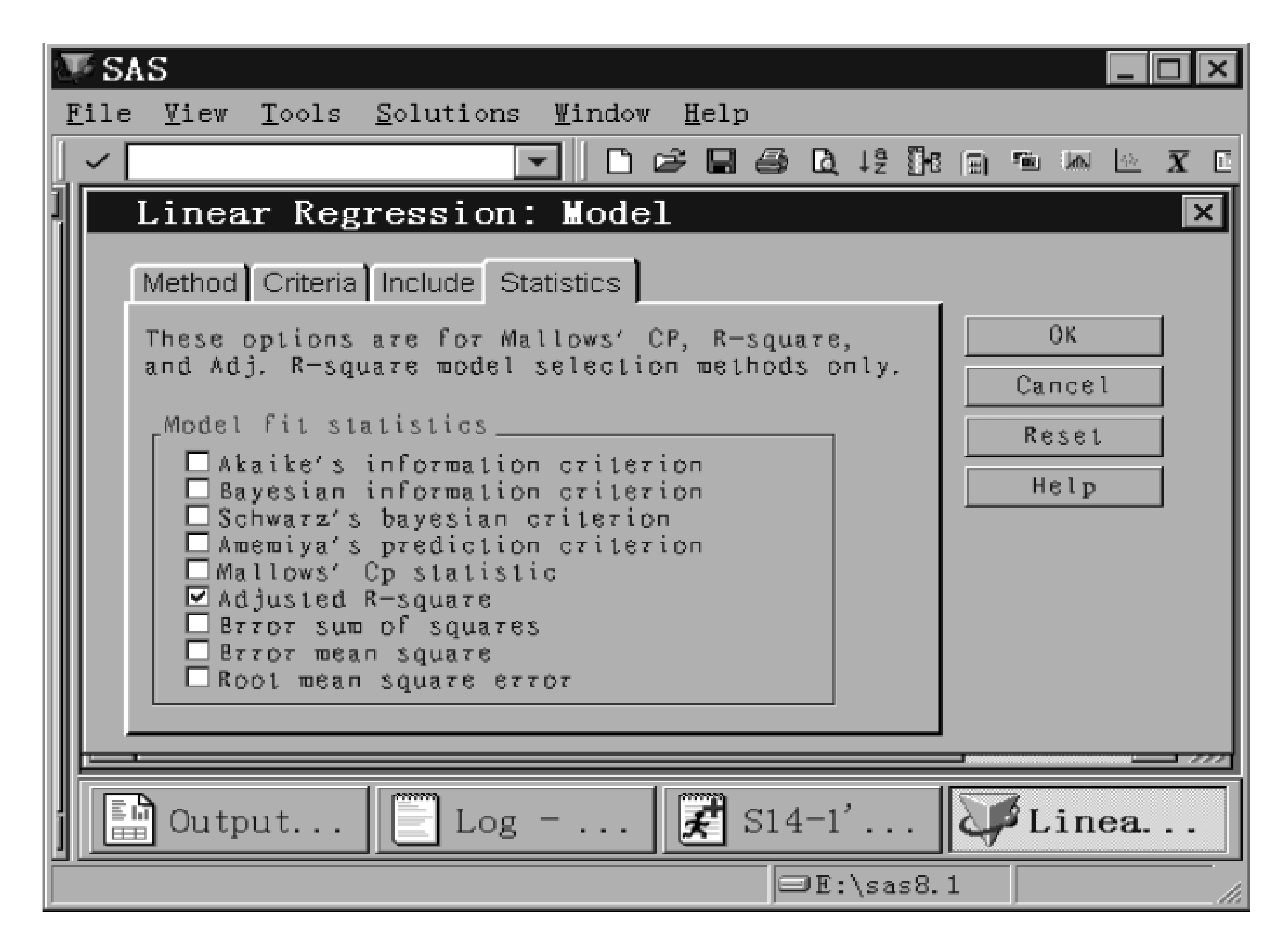


图 11.7 选择 Adjusted R-square 法

(9) 单击 OK 按钮,再单击 Statistics 标签,进入图 11.8,并选择平方和的 Type Ⅰ 和 Type Ⅱ 等(见带"√"项)。

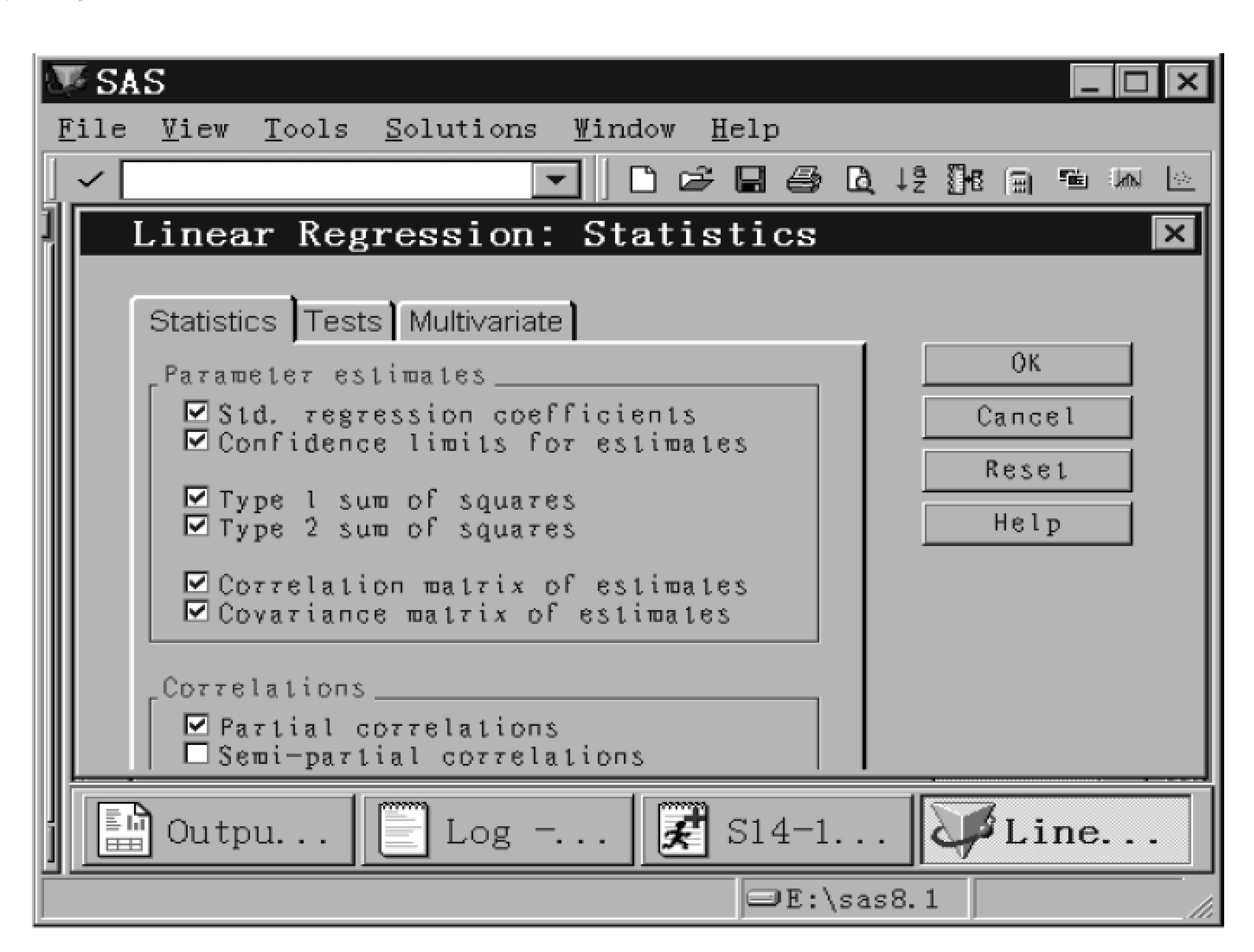


图 11.8 选择平方和的类型 Ⅰ 和 Ⅱ 等

(10) 单击 OK 按钮,再单击 Predictions,进入图 11.9,并选择统计量,见带"√"项。

(11) 单击两次 OK 按钮,输出图 11.10 的结果。

2. 回归分析

图 11.10 的回归分析见下面的 11.3 节。

下一节采用程序语句对程序 11.1 中的数据进行回归分析。

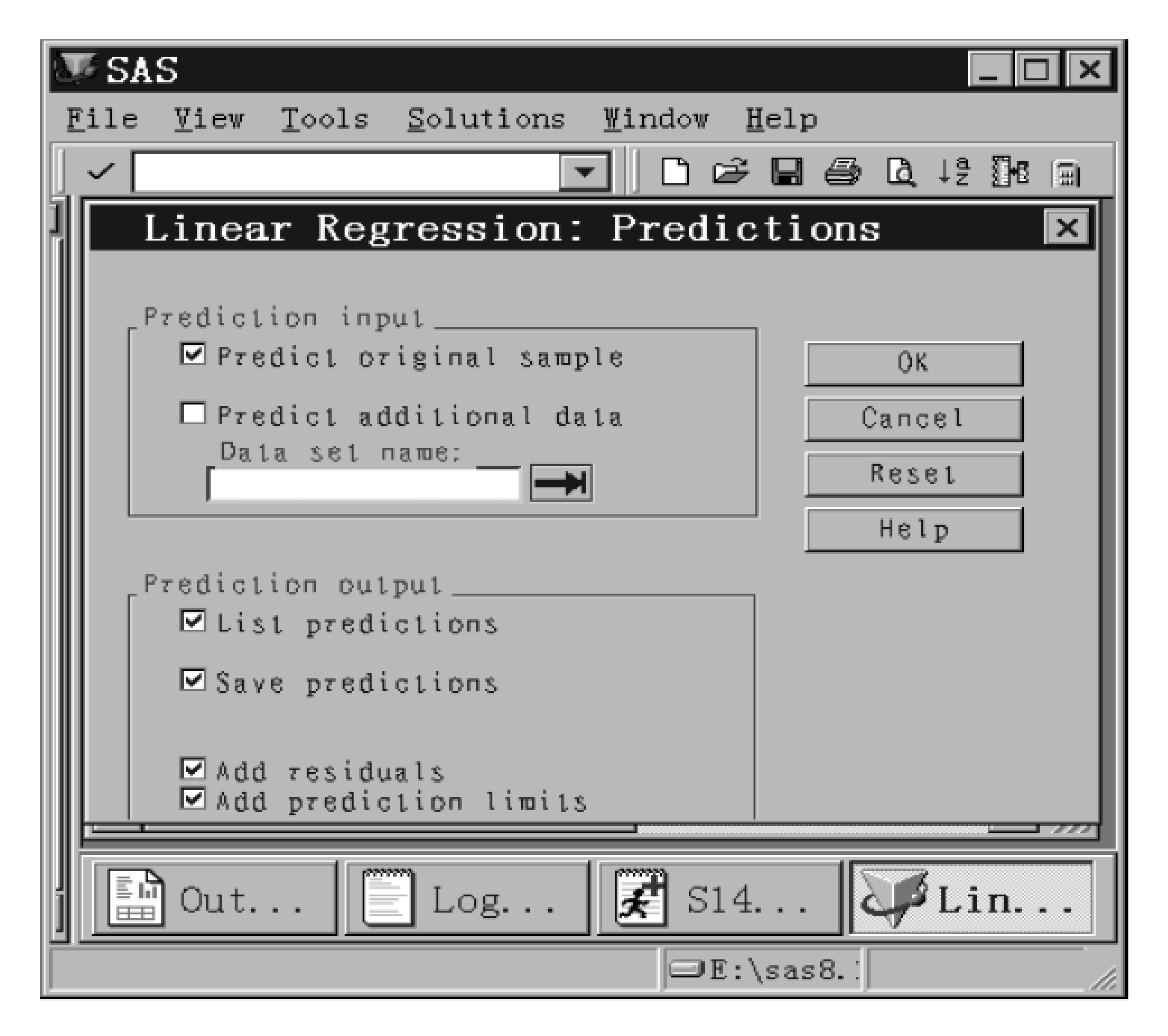
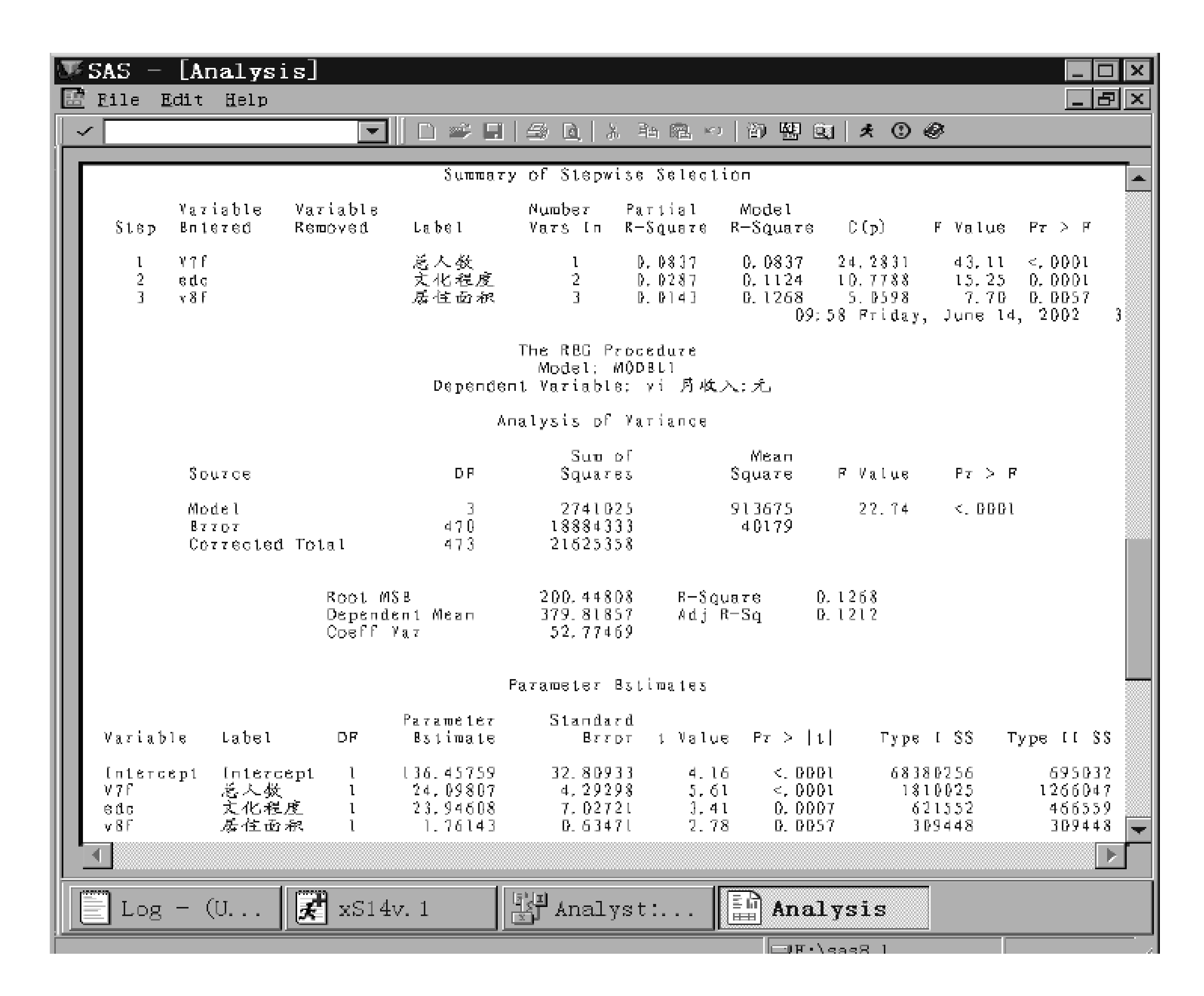


图 11.9 选择带“√”项的统计量



Summary of Stepwise Selection

Step	Variable Entered	Variable Removed	Label	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	V7f		总人数	1	0.0837	0.0837	24.2831	43.11	<.0001
2	edc		文化程度	2	0.0287	0.1124	10.7788	15.25	0.0001
3	v8f		居住面积	3	0.0143	0.1268	5.0598	7.70	0.0057

09:58 Friday, June 14, 2002 3

The REG Procedure
Model: MODEL1
Dependent Variable: vi 月收入:元

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	2741025	913675	22.74	<.0001
Error	470	18884333	40179		
Corrected Total	473	21625358			

Root MSE	200.44808	R-Square	0.1268
Dependent Mean	379.81857	Adj R-Sq	0.1212
Coeff Var	52.77469		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS	Type II SS
Intercept	Intercept	1	136.45759	32.80933	4.16	<.0001	68380256	695032
V7f	总人数	1	24.09807	4.29298	5.61	<.0001	1810025	1266047
edc	文化程度	1	23.94608	7.02721	3.41	0.0007	621552	466559
v8f	居住面积	1	1.76143	0.63471	2.78	0.0057	309448	309448

图 11.10 采用对话框法输出的结果示意图

11.2 REG过程的语句格式

本节采用程序 11.2 最后 5 行语句，替代图 11.2 至图 11.9 各个对话框的操作，不但同样能产生图 11.10 的回归输出，而且统计效果更佳（详见图 11.12 至图 11.25 所示的结果）。

程序 11.2：

```
DATA SQ14;
```

```
INFILE'd:\usersas\S474.dat';
INPUT Id 1-2 CASEID 3-5 AGE 6-7 SEX 8 edc 9 wk 10 fm 11 V6f 12 V7f 13-14
      v8f 15-16 V9f 17 V10A 18 V10B 19 V10C 20 V10D 21 vi 22-24 vo 25-27;
      oi=vo/vi; VIO=VI-VO; AV8F=V8F/V7F;av=vi/v7f;
LABEL SEX='户主性别' edc='文化程度' wk='具体工作' ID='区与街道代号'
      CASEID='问卷号' fm='婚姻状况' v6f='几代人?' V7f='总人数' V8f='居住面积'
      V9f='住房类型' V10A='煤气' V10B='卫生间' V10C='暖气' V10d='自来水'
      Vi='月收入:元' Vo='月开支:元';
PROC FORMAT;
PROC CORR;
  VAR VI V7F WK EDC SEX V8F;
PROC REG ;
  MODEL VI=V7F WK EDC SEX V8F /METHOD=STEPWISE SLE=.05 SLS=.05;
  MODEL VI=V7F WK EDC SEX V8F/METHOD=F SLE=.05;
  MODEL VI=V7F WK EDC SEX V8F/METHOD=B SLE=.05;
RUN;
```

运行程序 11.2 将产生下面图 11.12 至图 11.22 所示的结果。

程序 11.2 中 PROC REG 过程以下各条命令及其格式详见 11.2.1 节。

11.2.1 REG 程序中的语句及任选项

REG 程序中的全部语句及其任选项(关键词),如图 11.11 所示。

```
/* 回归过程的命令格式 */;

   PROC  REG  DATA=数据集名  OUTEST=名2  OUTSSCP=名3  NOPRINT  SIMPLE
         USSCP ALL  COVOUT  CORR  SINGULAR=N;
     MODEL  因变量=回归自变量v1  v2 ...
         METHOD=NONE|FORWARD|BACKWARD|STEPWISE|MAXR|MINR|RSQUARE
         SLEntry=值1  SLStay=值2  SELECT=...  INCLUDE=...  START=M
            STOP=N  NOPRINT  NOINT  ALL  XPX  I SS1  SS2
            STB  P  R  CLI  CLM  VIF  COVB  CORRB  COLLIN
            COLLINOINT TOL  DW  INFLUENCE  PARTIAL  DETAILS
            SIGAM=值  ADJRSQ  AIC  BIC  CP  GMSEP  JP  MSE  PC
            RMSE  SBC  SP  SSE  B;
      VAR  变量表;
      ID   变量;
      FRE  变量;
      WEIGHT  变量;
      ADD   变量表
      DELETE  变量表;
      DELOBS  n;
      PRINT ALL  XPX  I  SS1  SS2  STB  TOL  VIF  COVB  CORRB  COLLIN
      COLLINOINT  P  R  CLM CLI  DW  INFLUENCE  PARTIAL ANOVA ;
      OUTPUT  OUT=数据集名  PREDICTED=  RESIDUAL=  L95M=值  U95M=值
         L95=值  U95=值  STDP=值  STDI=值  STDR=值  STUDENT=值
         COOKD=值 H=值  PRESS=值  RSTUDENT=值  DFFITS=值  COVRATIO=值;
      TEST  式1,式2,...式k /PRINT;
      MTEST|MTEST  1,2,...k/PRINT  CANPRINT  DETAILS;
      BY  变量表;
```

图 11.11 Regression 的命令格式(见 A 盘的 S14.0 文件)

11.2.2 REG 程序中主要语句及关键词的注解

1. 下面是写在“PROC REG”后面的关键词

(1) DATA=数据集名:此项不可省(见图 11.11)。

(2) OUTEST＝名称：存储输出的估计参数，可省。

(3) OUTSSCP＝名称：存储输出的 SSCP 矩阵，可省。

(4) CORR：要求显示相关系数。

2. 写在 MODEL 之后的关键词

(1) Dependents：指定回归模型中的因变量名。

(2) Regressors：指定回归模型中的自变量名。

(3) /METHOD＝：指定回归分析的某种方法。其中，NONE 为取消各种回归；MAXR 要求按照最大 R^2 改善技术来选取最佳模型。

(4) SLENTRY＝α 值：默认为 $\alpha=0.05$。它是变量入选时的显著性水平：主要用于向前(F)和逐步回归中。

(5) SLSTAY＝α 值：默认为 0.05。它是变量被淘汰时的显著性水平；主要用于向后(B)和逐步回归中。

(6) SELECT＝自变量名。

(7) START＝值：指定最少估计多少个自变量。

(8) STOP＝值：指定最多估计多少个自变量。

(9) I：显示$(X'X)^{-1}$矩阵。若选择 I，则在图 11.13 中，将显示“X'X Inverse，B，SSE”标题。

(10) SS1：显示 TYPE Ⅰ SS 的顺序平方和，详见图 11.12 中的 Parameter Estimates 栏目下面的 SS1 值(806951)。

(11) SS2：显示模型参数偏平方和的估计值，即 TYPE Ⅱ SS。详见图 11.12 中② Parameter Estimates 栏目下面的 SS2 值(806951)。

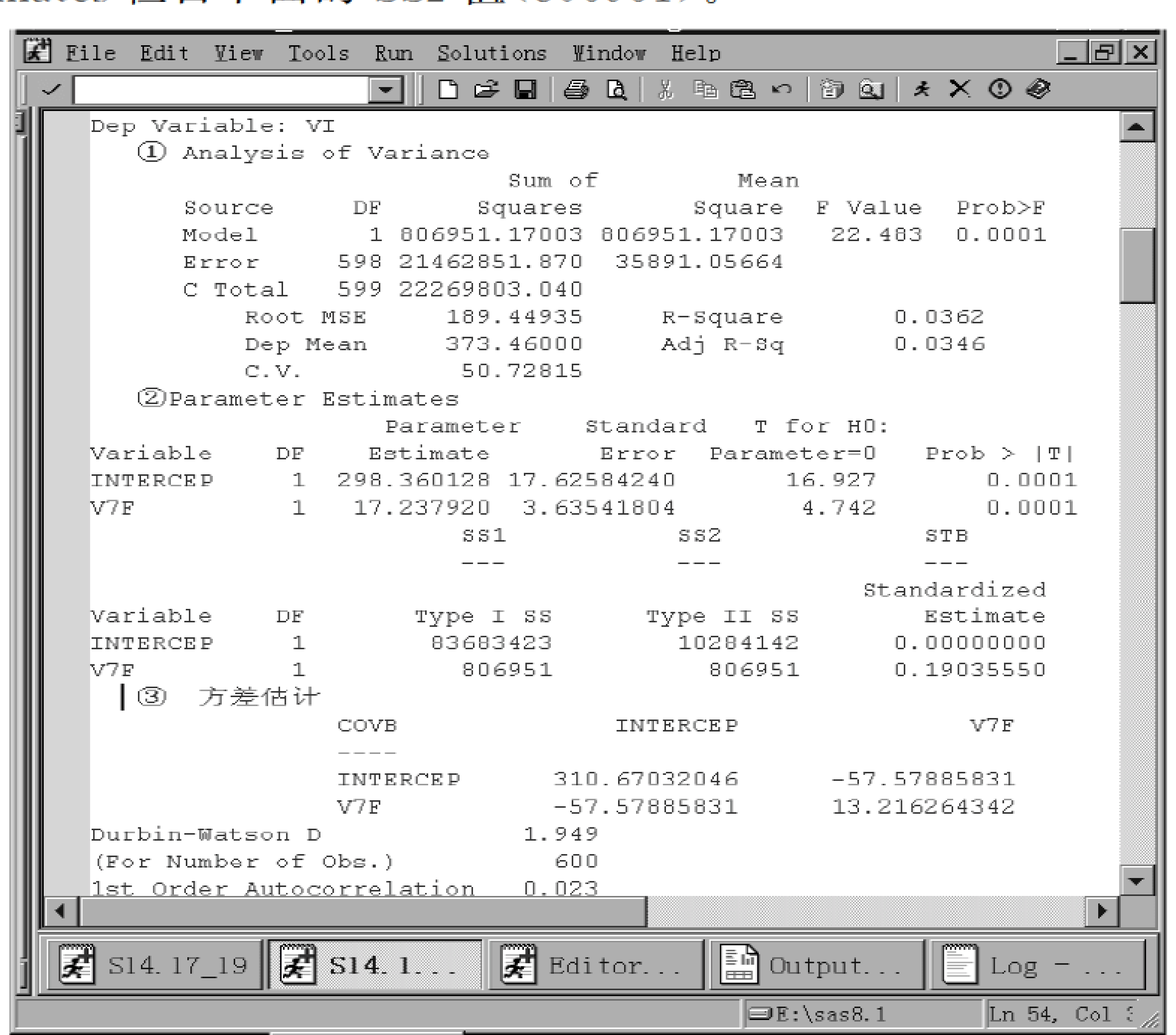


图 11.12 MODEL Vi＝v7f 的回归分析示意图

```
                    ④ Model Crossproducts X'X X'Y Y'Y
                      ------------------------------

X'X              INTERCEP             V7F
INTERCEP              600            2614   Intercept
V7F                  2614           14104   总人数
WK                   1854            8008   具体工作
VI                 224076         1023037
                    ⑤ Model Crossproducts X'X X'Y Y'Y
                      ------------------------------

X'X                    WK              VI
INTERCEP             1854          224076   Intercept
V7F                  8008         1023037   总人数
WK                   8608          699648   具体工作
VI                 699648       105953226

                ⑥X'X Inverse, Parameter Estimates, and SSE
                  ----------------------------------------

INVERSE            INTERCEP             V7F              WK           VI
INTERCEP       .01221719955   -.0016326409   -.0011125116  288.96267743
V7F            -.0016326409   .00036845886 8.8635705E-6    17.31279122
WK             -.0011125116   8.8635705E-6 .00034753916  2.9356835444
VI             288.96267743    17.31279122 2.9356835444    21438053.98
                         ⑦ Analysis of Variance
                           --------------------

                           Sum of          Mean
      Source     DF       Squares        Square   F Value      Prob>F
      Model       2  831749.06022  415874.53011    11.581      0.0001
      Error     597  21438053.980   35909.63816
      C Total   599  22269803.040
          Root MSE      189.49839     R-Square        0.0373
          Dep Mean      373.46000     Adj R-Sq        0.0341
          C.V.           50.74128
```

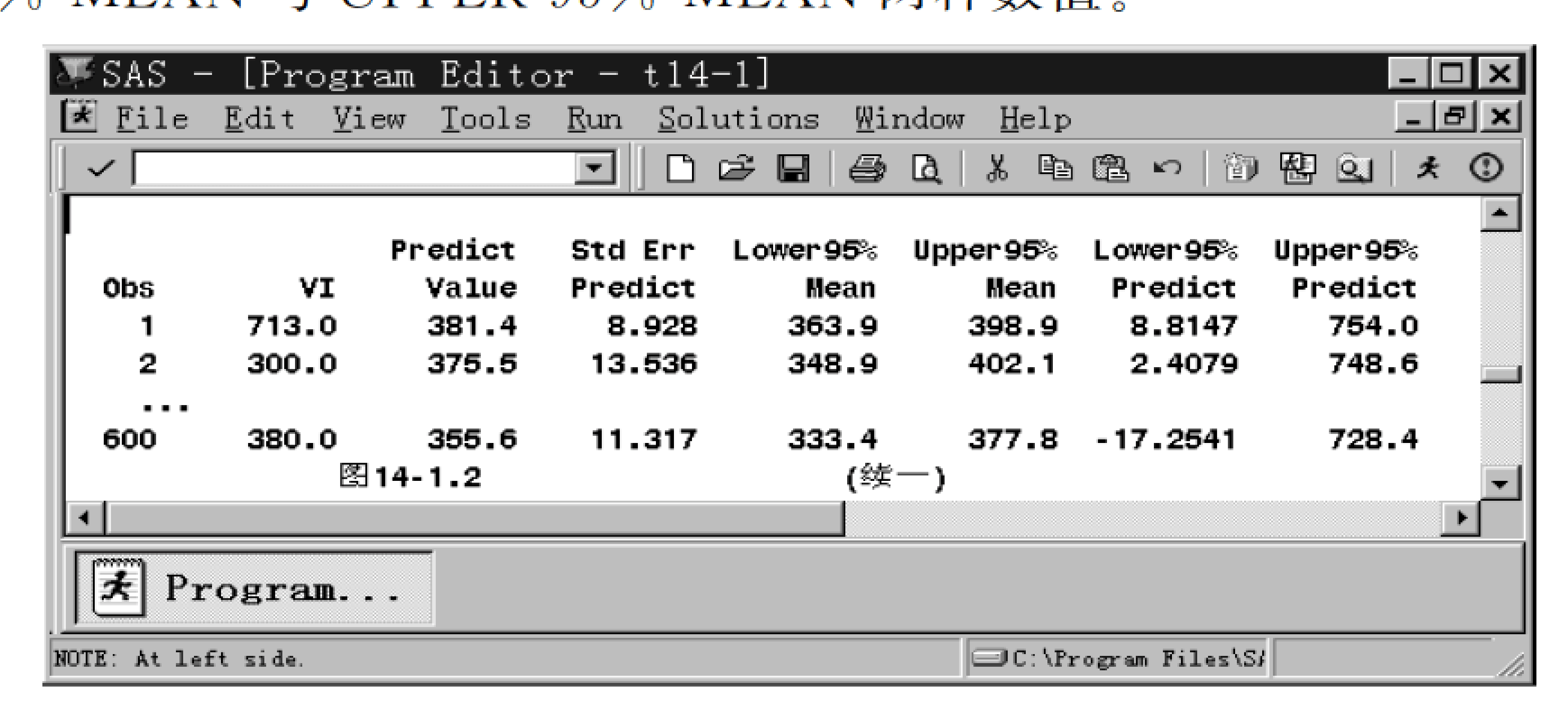
图 11.13　由 MODEL Vi=v7f wk/all dw influence 语句产生的回归图(待续)

(12) STB：显示标准化的回归系数，其值=估计的回归系数÷样本的标准误差=0.19035550。

(13) COVB：显示估计的协方差矩阵。该矩阵=$(X'X)^{-1} * MSE$。其中，MSE 为均方误差。详见图 11.12 的 Covariance Of Estimates 栏或 COVB 栏(矩阵)。

(14) Residual：表示每个观察值分析后的残差。其值=实际值-期望值。参阅图 11.15(b)。

(15) CLM：显示每一个因变量 Y 的期望值的 95%上下限置信度。详见图 11.14 的 LOWER 95% MEAN 与 UPPER 95% MEAN 两种数值。



Obs	VI	Predict Value	Std Err Predict	Lower95% Mean	Upper95% Mean	Lower95% Predict	Upper95% Predict
1	713.0	381.4	8.928	363.9	398.9	8.8147	754.0
2	300.0	375.5	13.536	348.9	402.1	2.4079	748.6
...							
600	380.0	355.6	11.317	333.4	377.8	-17.2541	728.4

图14-1.2　(续一)

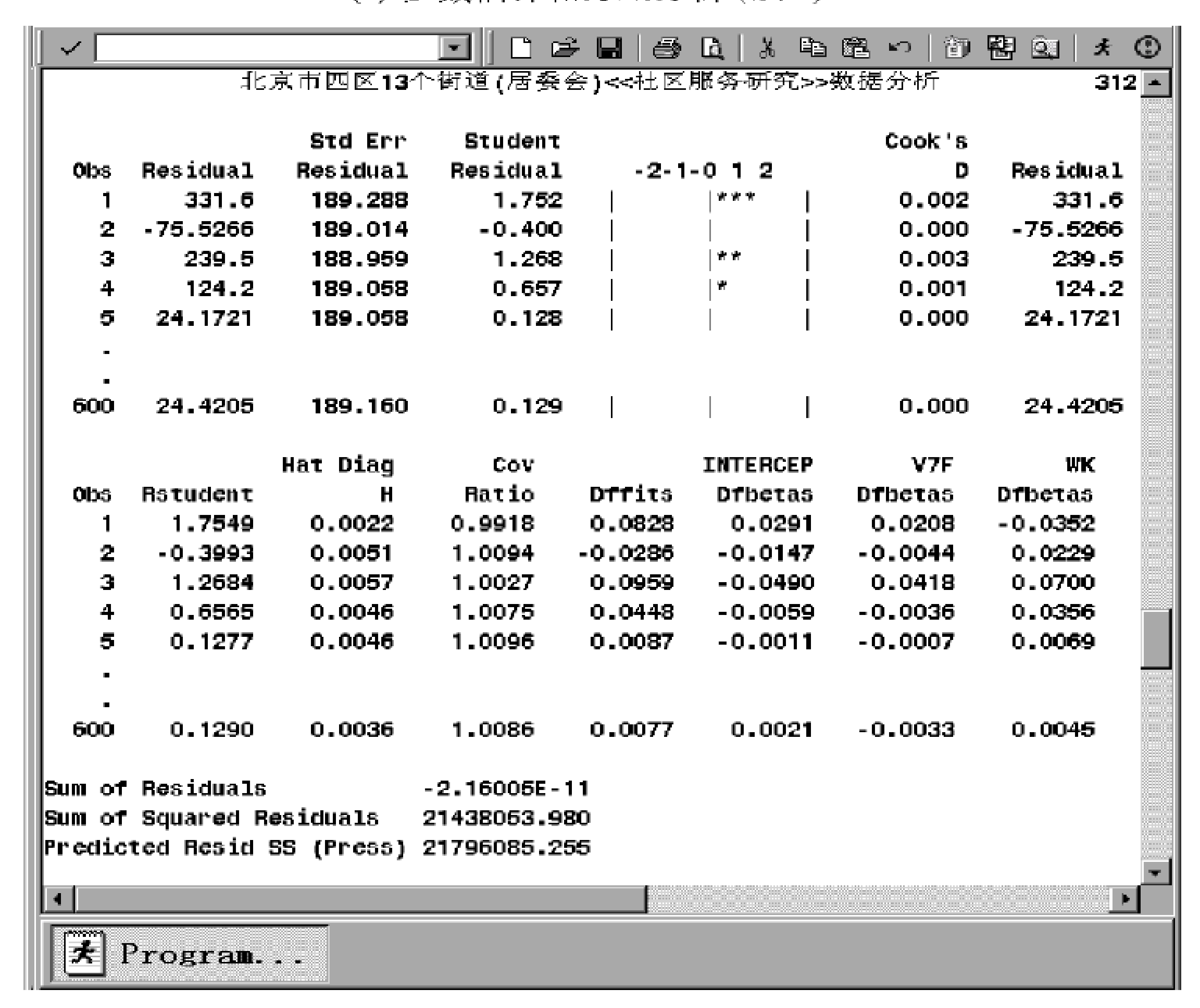
图 11.14　由 MODEL Vi=v7f wk/all dw influence 语句产生的回归图(续 1)

(16) CLI：显示各个观察值的95%上限与95%下限置信度，详见图11.14。

(17) DW：显示Durbin-Watson D统计量，供检验是否有第一阶自我相关。它适合时序分析，特别适合于前后期有相互影响的数据。可参阅图11.12中的Durbin-Watson D行上的相应值，DW=1.949，而且自我相关系数γ=0.023，小而合格。说明两点间的残差互为独立。

```
    ⑧ Parameter Estimates
                Parameter      Standard     T for H0:
Variable DF      Estimate         Error   Parameter=0    Prob > |T|
INTERCEP  1    288.962677   20.94552971        13.796        0.0001
V7F       1     17.312791    3.63747498         4.760        0.0001
WK        1      2.935684    3.53270511         0.831        0.4063
                                         Standardized
Variable DF     Type I SS   Type II SS       Estimate     Tolerance
INTERCEP  1      83683423      6834580     0.00000000      .
V7F       1        806951       813477     0.19118229    0.99938649
WK        1         24798        24798     0.03337970    0.99938649
                            Variance
Variable     DF            Inflation
INTERCEP      1           0.00000000
V7F           1           1.00061389
WK            1           1.00061389
    ⑨ Covariance of Estimates
  COVE                INTERCEP               V7F                  WK
  INTERCEP        438.71521497      -58.62754219        -39.94988972
  V7F             -58.62754219      13.231224211        .31828760979
  WK              -39.94988972      .31828760979        12.480005403
    ⑩ Correlation of Estimates
  CORRB               INTERCEP               V7F                  WK
  INTERCEP              1.0000           -0.7695             -0.5399
  V7F                  -0.7695            1.0000              0.0248
  WK                   -0.5399            0.0248              1.0000
Durbin-Watson D                1.945
(For Number of Obs.)             600
1st Order Autocorrelation      0.025
```

(a) 参数估计和方差分析(续2)



北京市西区13个街道(居委会)<<社区服务研究>>数据分析 312

Obs	Residual	Std Err Residual	Student Residual	-2-1-0 1 2	Cook's D	Residual
1	331.6	189.288	1.752	| |*** |	0.002	331.6
2	-75.5266	189.014	-0.400	| | |	0.000	-75.5266
3	239.5	188.959	1.268	| |** |	0.003	239.5
4	124.2	189.058	0.657	| |* |	0.001	124.2
5	24.1721	189.058	0.128	| | |	0.000	24.1721
.						
.						
600	24.4205	189.160	0.129	| | |	0.000	24.4205

Obs	Rstudent	Hat Diag H	Cov Ratio	Dffits	INTERCEP Dfbetas	V7F Dfbetas	WK Dfbetas
1	1.7549	0.0022	0.9918	0.0828	0.0291	0.0208	-0.0352
2	-0.3993	0.0051	1.0094	-0.0286	-0.0147	-0.0044	0.0229
3	1.2684	0.0057	1.0027	0.0959	-0.0490	0.0418	0.0700
4	0.6565	0.0046	1.0075	0.0448	-0.0059	-0.0036	0.0356
5	0.1277	0.0046	1.0096	0.0087	-0.0011	-0.0007	0.0069
.							
.							
600	0.1290	0.0036	1.0086	0.0077	0.0021	-0.0033	0.0045

Sum of Residuals -2.16006E-11
Sum of Squared Residuals 21438053.980
Predicted Resid SS (Press) 21796085.255

Program...

(b) Cook距离和残差分析(续3)

图 11.15

(18) INFLUENCE：要求 SAS 对每个观察值的估计值与预测值作更详细的相互影响的分析。

(19) ALL：要求 SAS 分析以下参数：

XPX，SS1，SS2，STB，TOL，COVB. CORRB，SEQB，P，R，CLM，CLI，SPEC，ACOV，PCORR1，PCORR2，SCORR1，SCORR2。详见图 11.13 至图 11.15 所示的结果。

(20) PARTIAL：要求 SAS 画出图 11.16 所示的每个回归自变量的"偏回归残差图"。

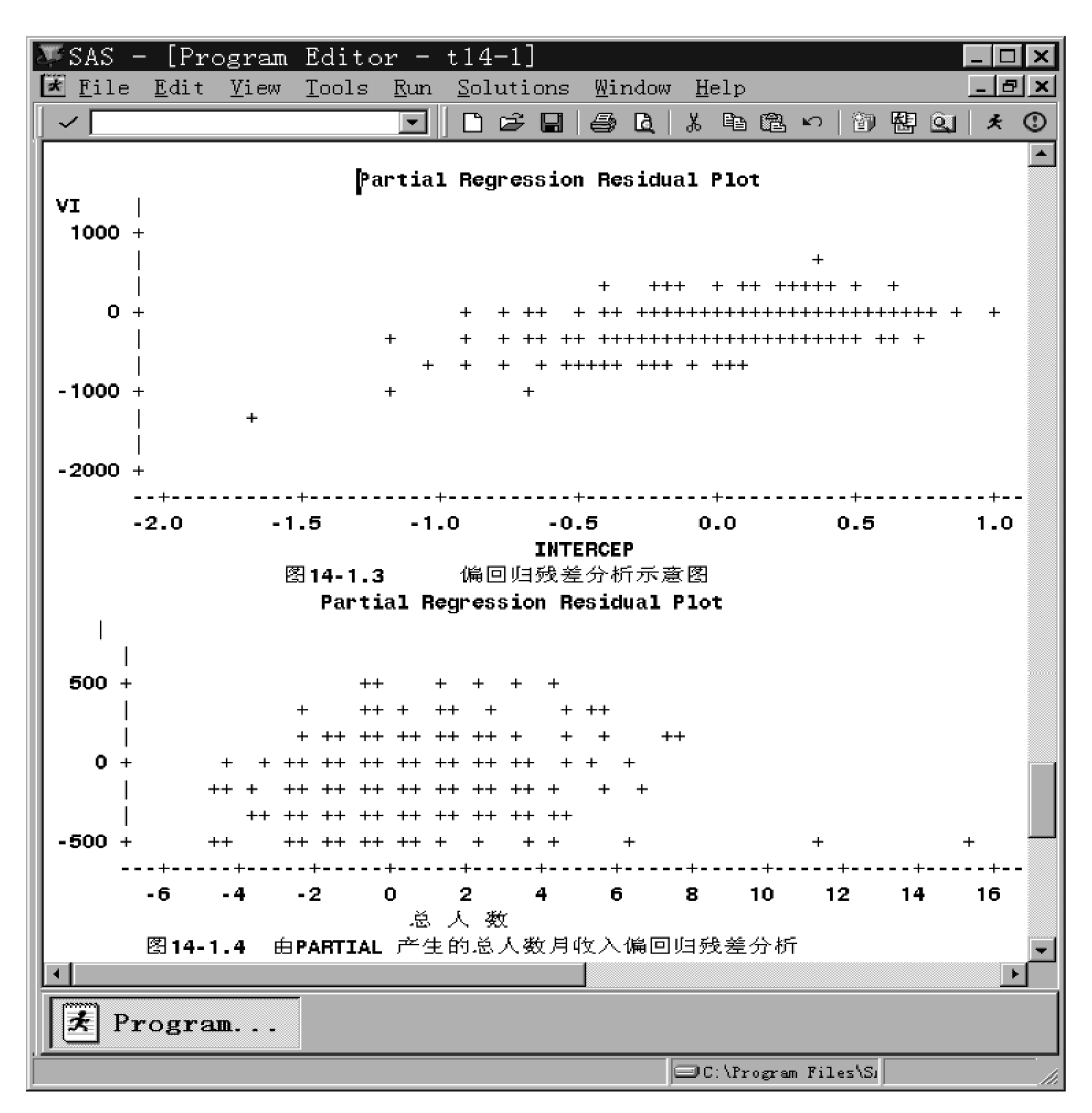


图 11.16 偏回归残差分析

(21) P=PRED：要求显示标题 PRED 表示"预测值"。

(22) L95=L95：显示结果中各个观察值下限的 95%置信区间。

U95=U95：显示各个观察值上限的 95%置信区间。

(23) R=RESID：显示残差值。

(24) COOKD=COOKD：显示 COOK 的 D 统计量，见图 11.15(b)。

(25) H=H：显示 $X_i(X'X)^{-1} * X_i$ 值，见图 11.13 第 1 个小标题。

11.3 REG程序进一步实例

本节通过程序 11.2 的例子，深入剖析逐步回归分析法、向前选择变量法和自后淘汰变量法的执行步骤及输出结果。

程序 11.3(如图 11.17 所示)：

图 11.17(程序 11.3)前面的 DATA、INFILE、INPUT 等语句用以定义数据的来源、数据的格式和栏目位置，IF…THEN…语句用以排除缺少值，LABEL 语句用以定义变量标签，VALUE 语句用以定义数值标签。

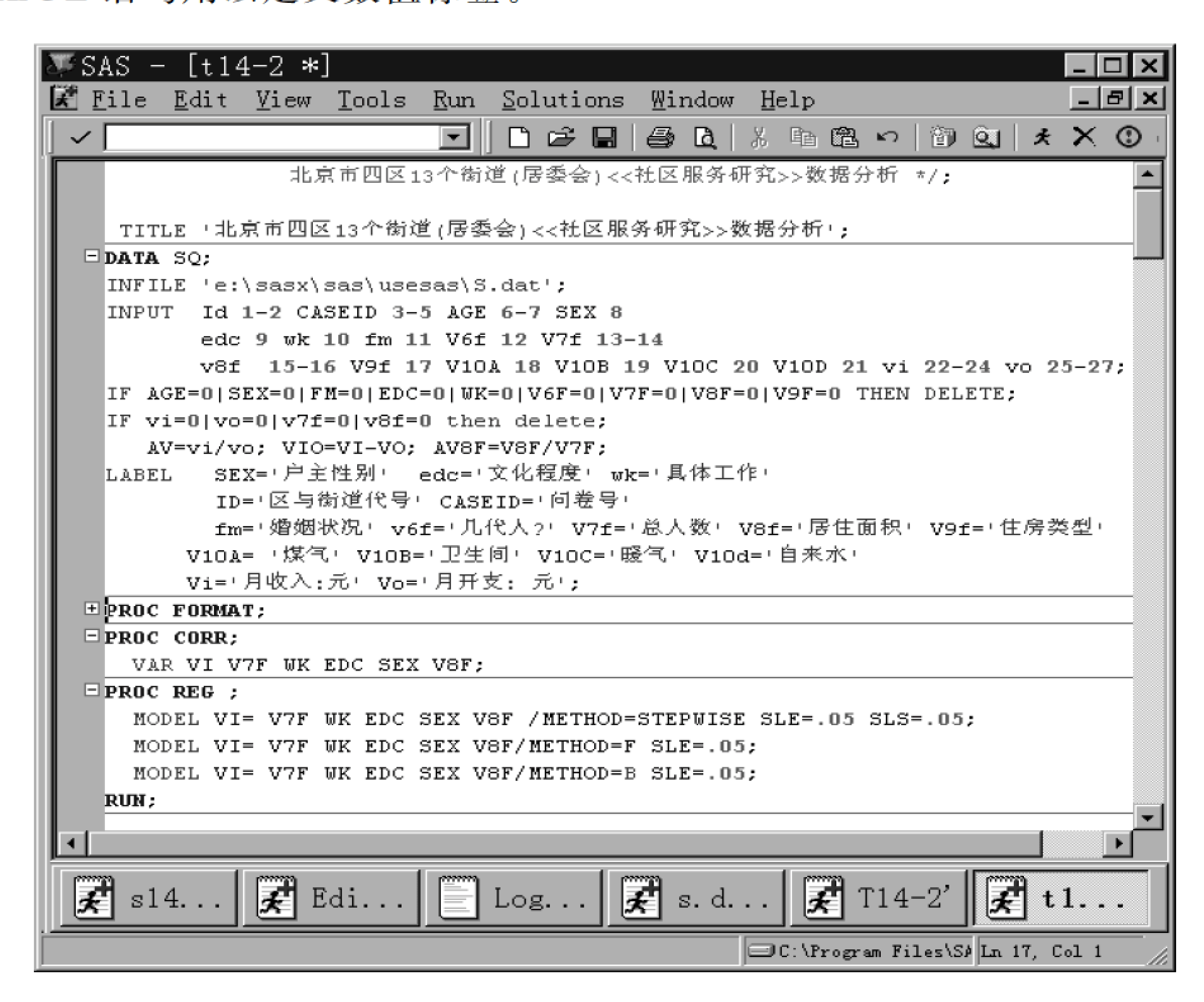


图 11.17 逐步回归分析法、向前选择变量法和自后淘汰变量法(程序 11.3)

程序 11.3 的 PROC CORR 语句用以计算相关矩阵。程序 11.3 的最后语句则是调用“PROC REG;”过程分别作逐步回归、向前选择变量以及自后淘汰变量的回归分析。

运行程序 11.3 产生图 11.18 至图 11.22 所示的结果。

下面将程序 11.3 中 PROC CORR 后面七行语句的执行结果，以图形方式逐一加以说明。

1. 程序 11.3 的 PROC CORR 主语句

由于该语句后面的 VAR 关键词中指明了当前月收入(VI)、家庭总人数(V7F)、性别(SEX)、教育水平(EDC)、工种(WK)和居住面积(V8F)共 6 个变量，所以程序运行到该

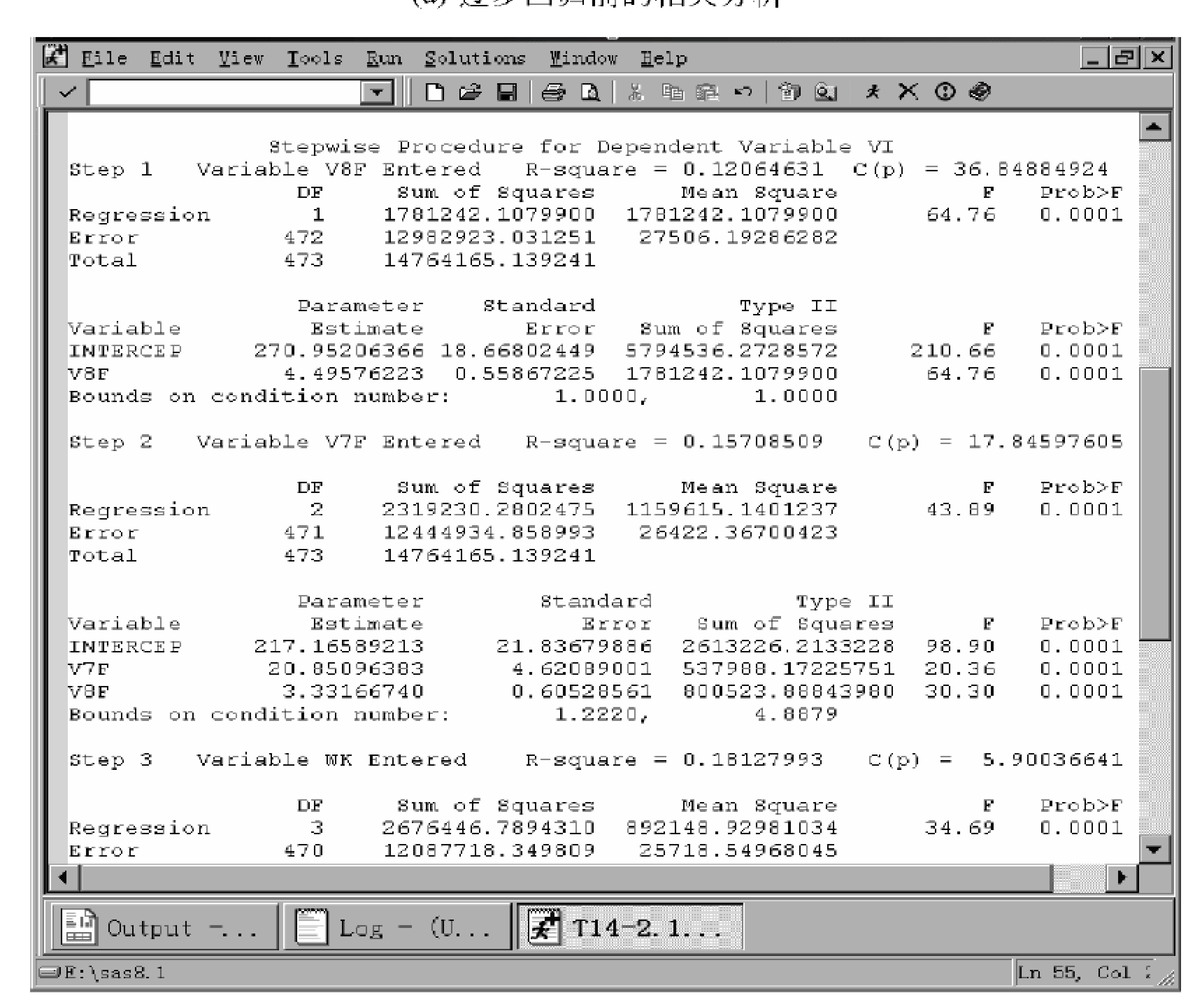
语句时，产生了如图 11.18(a)所示的相关分析示意图。图 11.18(a)的上半图为默认的而且是最基本的统计量，即：观察值 N、各变量的均值 Mean、标准偏差 Std Dev、各变量值之和 Sum 及各变量的最小值最大值。

```
             北京市四区13个街道(居委会)<<社区服务研究>>数据分析            74

Variable     N        Mean     Std Dev          Sum     Minimum     Maximum
VI         474   408.10127   176.67450       193440    50.00000         999
V7F        474     4.28270     1.78798         2030     1.00000    12.00000
WK         474     3.44937     1.99962         1635     1.00000     6.00000
EDC        474     3.42616     1.28737         1624     1.00000     5.00000
SEX        474     1.51899     0.50017    720.00000     1.00000     2.00000
V8F        474    30.50633    13.64984        14460     7.00000    99.00000

      Correlation Coefficients / Prob > |R| under Ho: Rho=0 / N = 474
                VI         V7F          WK         EDC         SEX         V8F
 VI        1.00000     0.32072    -0.13462     0.06779    -0.04010     0.34734
            0.0000      0.0001      0.0033      0.1406      0.3838      0.0001
 V7F       0.32072     1.00000    -0.00900    -0.12960     0.07909     0.42621
            0.0001      0.0000      0.8451      0.0047      0.0854      0.0001
 WK       -0.13462    -0.00900     1.00000     0.08888     0.08129     0.08568
            0.0033      0.8451      0.0000      0.0531      0.0771      0.0623
 EDC       0.06779    -0.12960     0.08888     1.00000    -0.18004     0.10861
            0.1406      0.0047      0.0531      0.0000      0.0001      0.0180
 SEX      -0.04010     0.07909     0.08129    -0.18004     1.00000    -0.04662
            0.3838      0.0854      0.0771      0.0001      0.0000      0.3111
 V8F       0.34734     0.42621     0.08568     0.10861    -0.04662     1.00000
            0.0001      0.0001      0.0623      0.0180      0.3111      0.0000
```

(a) 逐步回归前的相关分析



```
File  Edit  View  Tools  Run  Solutions  Window  Help

                Stepwise Procedure for Dependent Variable VI
Step 1   Variable V8F Entered   R-square = 0.12064631  C(p) = 36.84884924
                DF      Sum of Squares       Mean Square          F   Prob>F
Regression       1     1781242.1079900   1781242.1079900      64.76   0.0001
Error          472     12982923.031251    27506.19286282
Total          473     14764165.139241

                Parameter    Standard           Type II
Variable         Estimate       Error    Sum of Squares          F   Prob>F
INTERCEP     270.95206366 18.66802449   5794536.2728572     210.66   0.0001
V8F            4.49576223  0.55867225   1781242.1079900      64.76   0.0001
Bounds on condition number:        1.0000,          1.0000

Step 2   Variable V7F Entered   R-square = 0.15708509   C(p) = 17.84597605

                DF      Sum of Squares       Mean Square          F   Prob>F
Regression       2     2319230.2802475   1159615.1401237      43.89   0.0001
Error          471     12444934.858993    26422.36700423
Total          473     14764165.139241

                Parameter          Standard            Type II
Variable         Estimate             Error     Sum of Squares       F   Prob>F
INTERCEP     217.16589213       21.83679886    2613226.2133228   98.90   0.0001
V7F           20.85096383        4.62089001    537988.17225751   20.36   0.0001
V8F            3.33166740        0.60528561    800523.88843980   30.30   0.0001
Bounds on condition number:        1.2220,          4.8879

Step 3   Variable WK Entered    R-square = 0.18127993   C(p) =  5.90036641

                DF      Sum of Squares       Mean Square          F   Prob>F
Regression       3     2676446.7894310   892148.92981034      34.69   0.0001
Error          470     12087718.349809    25718.54968045

Output -...   Log - (U...   T14-2.1...
E:\sas8.1                                                        Ln 55, Col :
```

(b) 逐步回归法的输出结果

图　11.18

图 11.18(a)的下半图为 6 个变量的相关矩阵，每对变量有两行系数，例如，Vi 与 V7f 该对变量的相关系数为 0.32072。又因为 0.32072 下面的 0.0001<0.05。说明月收入与家庭总人数确有一定的关系。

2. 程序 11.3 的"MODEL Vi＝V7F WK…/METHOD＝STEPWISE…"语句

该语句要求以 Vi 为因变量，以 V7F WK 等 5 个变量为自变量进行逐步回归分析，详见图 11.18(b)。

(1) Step 1：这是逐步回归的第 1 步，第 1 个进入回归模型的变量是 V8F(居住面积)，其回归平方和为 1781242.10799，总平方和 SST＝14764165.139241，F 值＝64.76。(注：F 值＝回归均方 MSR÷误差均方 MSE＝1781242.10799÷27506.19286282＝64.76)。

因为 P 值＝0.0001＜α＝0.05，因此 V8F 变量入选到模型中。或说 $F_{0.05}(1,472)$＝3.84，64.76 值＞3.84($F_{0.05}$值)，因此 V8F 变量入选到模型中。

(2) Step 2：这是逐步回归的第 2 步，即：决定第 2 个入选的自变量。此时可由 TYPE Ⅱ SS 的值判定在 V8F 入选后，哪一个变量对回归模型的贡献较大。显见，继 V8F 之后，对模型贡献较大的变量为 V7F，而且其 F 值＝20.36，它大于 $F_{0.05}(1,471)$＝3.84，因此 V7F 也应入选。另一方面，最先入选的 V8F 变量的 TYPE Ⅱ SS 值为 800523.8884398＞537988.1722751，而且其 F 值为 30.30＞3.84，因此也不能排除 V8F 变量。换言之，第二个变量 V7F 入选后不排挤第一个变量的入选。

(3) Step 3：如图 11.19 所示，除了 V8F 和 V7F 入选有效外，其余变量中对模型贡献稍大的则是变量 WK，其 TYPE Ⅱ SS 值为 357216.50918355，而且 F 值为 13.89＞$F_{0.05}(1,470)$＝3.84，因此 WK 有资格入选。同时，第 3 个变量入选后，对前两个变量的参数重新计算之后，显见：V8F 的 F 值为 35.05＞$F_{0.05}(1,470)$＝3.84，因此 V8F 应继续留在模型内；V7F 的 F 值为 19.18＞$F_{0.05}(1,470)$＝3.84，因此 V7F 也应留在模型中。

此时，可挖掘出最佳回归模型应包含三个自变量：V8F，V7F 和 WK，因此，回归方程式为：

$$\begin{aligned}VI &= \beta_0 + \beta_1 * V7F + \beta_2 * WK + \beta_3 * V8F \\ &= 261.735 + 19.992 * V7F - 13.812 * WK + 3.553 * V8F\end{aligned}$$

(4) 命令中原有 5 个自变量，本应作五步回归，但进行到 step3 后，其余两个变量(EDC 和 SEX)由于显著性水平大于 0.05，因此中止回归。

(5) 图 11.19 的最后 3 行，有一个综合分析表，说明逐步回归的结果、回归模型中应包含的变量情形，在图 11.19 中显示的是 V8F，V7F 和 WK 三个变量。

3. 程序 11.3 的"MODEL VI＝V7F WK…/METHOD＝F…"语句

该语句要求以 VI 为因变量，以 V7F 、WK 等 5 个变量为自变量，进行 FORWARD(向前选择变量法)回归分析，详见图 11.20。

向前选择变量法的特点是：一旦变量入选，就无法从模型中删除，这是最大的缺点。因为，某些变量进入模型后，通过参数重新计算，有可能对模型不再产生显著的贡献。以

```
Total             473    14764165.139241

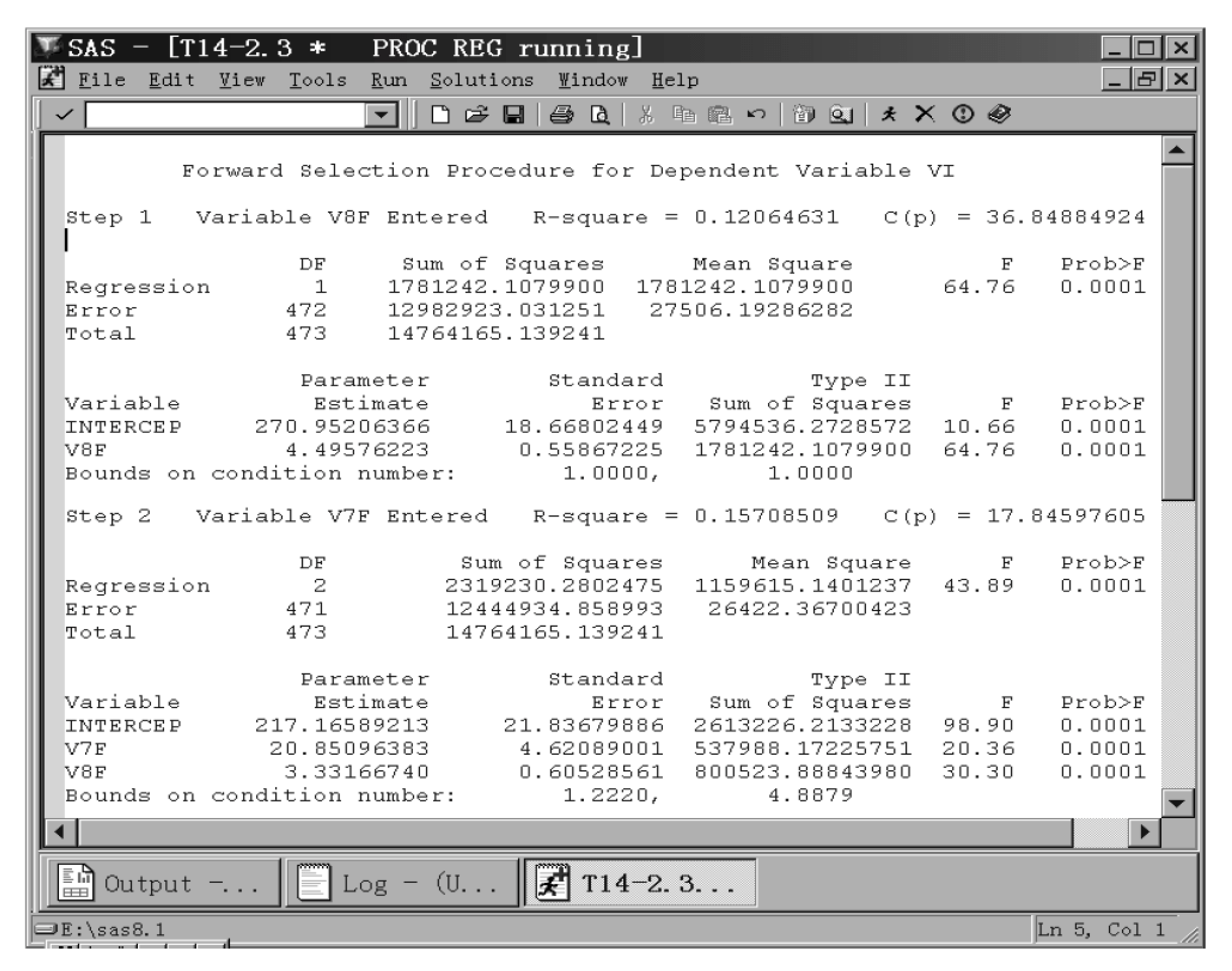
                   Parameter          Standard           Type II
Variable            Estimate             Error      Sum of Squares      F     Prob>F
INTERCEP        261.73456603       24.64054752  2901803.0307146 112.83    0.0001
V7F              19.99185587        4.56475513   493307.50972048  19.18    0.0001
WK              -13.81152816        3.70594705   357216.50918355  13.89    0.0002
V8F               3.55298733        0.60011515   901497.68248007  35.05    0.0001
Bounds on condition number:            1.2341,        10.4074

All variables in the model are significant at the 0.0500 level.
No other variables met the 0.0500 significance level for entry
into the model.

       Summary of Stepwise Procedure for Dependent Variable VI
       Variable         Number  Partial   Model
Step   Entered Removed  In       R**2      R**2       C(p)            F     Prob>F

   1    V8F               1     0.1206    0.1206    36.8488      64.7579    0.0001
   2    V7F               2     0.0364    0.1571    17.8460      20.3611    0.0001
   3    WK                3     0.0242    0.1813     5.9004      13.8894    0.0002
```

图 11.19　逐步回归法的输出结果



```
SAS - [T14-2.3 *    PROC REG running]
File  Edit  View  Tools  Run  Solutions  Window  Help

        Forward Selection Procedure for Dependent Variable VI

Step 1   Variable V8F Entered    R-square = 0.12064631    C(p) = 36.84884924

                   DF      Sum of Squares      Mean Square         F    Prob>F
Regression          1     1781242.1079900  1781242.1079900     64.76    0.0001
Error             472     12982923.031251   27506.19286282
Total             473     14764165.139241

                 Parameter         Standard           Type II
Variable          Estimate            Error     Sum of Squares        F    Prob>F
INTERCEP      270.95206366      18.66802449    5794536.2728572    10.66    0.0001
V8F             4.49576223       0.55867225    1781242.1079900    64.76    0.0001
Bounds on condition number:           1.0000,          1.0000

Step 2   Variable V7F Entered    R-square = 0.15708509    C(p) = 17.84597605

                   DF          Sum of Squares       Mean Square        F    Prob>F
Regression          2         2319230.2802475   1159615.1401237    43.89    0.0001
Error             471         12444934.858993    26422.36700423
Total             473         14764165.139241

                 Parameter         Standard           Type II
Variable          Estimate            Error     Sum of Squares        F    Prob>F
INTERCEP      217.16589213      21.83679886    2613226.2133228    98.90    0.0001
V7F            20.85096383       4.62089001    537988.17225751    20.36    0.0001
V8F             3.33166740       0.60528561    800523.88843980    30.30    0.0001
Bounds on condition number:           1.2220,          4.8879

Output -...   Log - (U...   T14-2.3...
E:\sas8.1                                                       Ln 5, Col 1
```

图 11.20　向前选择法的输出结果

图 11.20 为例。

Step 1：首先入选的变量是 V8F（居住面积）。其回归平方和（SSR）为 1781242.10799，F 值＝64.7＞$F_{0.05}$（1，472）＝3.84，这就是 V8F 入选到模型的原因。

Step 2：第 2 步其次入选的变量是 V7F（家庭总人数），其 F 值＝20.36＞$F_{0.05}$（1，471）＝3.84，也令人满意。

Step 3：如图 11.21 所示，第 3 步入选的变量是 WK（工种），其 F 值＝13.89＞$F_{0.05}$(1,470)＝3.84，该变量 WK 也应该入选。同时，前两步入选的变量，其 F 值保持大于 3.84，因此 3 个变量都有资格保持留在模型中。

```
Step 3   Variable WK Entered    R-square = 0.18127993    C(p) =  5.90036641

                  DF      Sum of Squares       Mean Square          F    Prob>F
Regression         3     2676446.7894310   892148.92981034      34.69    0.0001
Error            470     12087718.349809    25718.54968045
Total            473     14764165.139241
                  Parameter         Standard           Type II
Variable           Estimate            Error    Sum of Squares        F    Prob>F
INTERCEP       261.73456603      24.64054752   2901803.0307146 112.83    0.0001
V7F             19.99185587       4.56475513   493307.50972048  19.18    0.0001
WK             -13.81152816       3.70594705   357216.50918355  13.89    0.0002
V8F              3.55298733       0.60011515   901497.68248007  35.05    0.0001
Bounds on condition number:        1.2341,        10.4074
No other variables met the 0.0500 significance level for entry
into the model.

   Summary of Forward Selection Procedure for Dependent Variable VI
        Variable    Number   Partial    Model
Step    Entered       In       R**2      R**2        C(p)            F    Prob>F
   1    V8F            1     0.1206    0.1206     36.8488      64.7579    0.0001
   2    V7F            2     0.0364    0.1571     17.8460      20.3611    0.0001
   3    WK             3     0.0242    0.1813      5.9004      13.8894    0.0002
```

图 11.21　向前选择法的输出结果（续）

由于回归进入第 3 步后，其他变量入选时的显著性水平均达到 0.05，所以回归中止。这时，在模型中的变量有 V8F，V7F 和 WK（见图 11.21 的最后结论）。

回归方程式与逐步回归中的方程式完全一样，即：

$$VI＝261.735＋19.992 * V7F－13.812WK＋3.553 * V8F$$

注意：由于回归进行到第 3 步便中止，因此本例未遇到向前选择法的缺点。即：看不到模型中的该排除而未排除的变量。

4. 程序 11.3 的“MODEL VI＝V7F WK…/METHOD＝B”语句

这是自后淘汰变量法的语句，该语句产生图 11.22 所示的结果。

自后淘汰变量法是首先将所有变量都放入模型里，然后一次一次地淘汰那些 F 值的显著性水平（Prob＞F）大于 0.10 值（查表值）的变量，以图 11.22 为例。

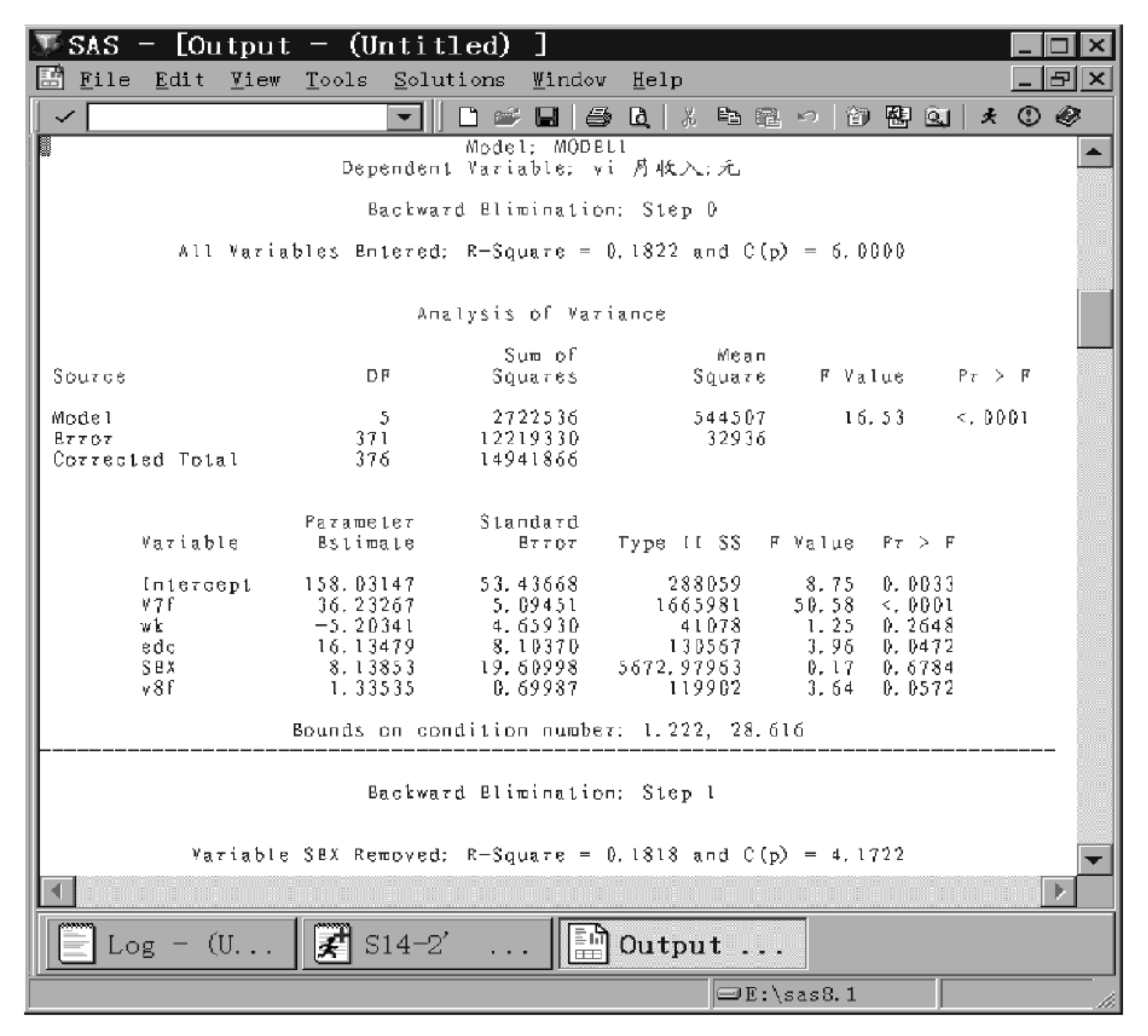
（1）Step 0：见图 11.22(a)第 3 行，该步是将 5 个变量都入选，它类似于完全回归模型。

（2）Step 1：但是从第 0 步看来，自变量 SEX 的 TYPE Ⅱ SS 值为最小（仅 5672.97963），且 F 值的显著性水平（Prob＞F）为 0.6784，大于 0.10 值，因此被淘汰。所以，在 Step 1 中，仅剩下 4 个自变量。

（3）在 Step 1 中，WK 变量的 TYPE Ⅱ SS 成为最小者（仅 38905），F 值的显著性水平（Prob＞F）为 0.2773 大于 0.10 值，因此，WK 变量也被淘汰。

（4）回归到 Step 1 时，由于 SEX、WK 变量淘汰时的显著性水平大于 0.10，因此中止淘汰。

最后得到回归方程式为：VI＝149.60289＋36.61874 * V7F＋16.95516 * EDC＋1.27267 * V8F。




```
                              Model: MODEL1
                  Dependent Variable: yi 月收入:元

                      Backward Elimination: Step 0

          All Variables Entered: R-Square = 0.1822 and C(p) = 6.0000


                          Analysis of Variance

                                   Sum of          Mean
Source                  DF        Squares        Square    F Value    Pr > F

Model                    5        2722536        544507      16.53    <.0001
Error                  371       12219330         32936
Corrected Total        376       14941866


                Parameter      Standard
      Variable   Estimate         Error    Type II SS  F Value  Pr > F

      Intercept  158.03147     53.43668       288059     8.75   0.0033
      v7f         36.23267      5.09451      1665981    50.58   <.0001
      wk          -5.20341      4.65930        41078     1.25   0.2648
      edc         16.13479      8.10370       130567     3.96   0.0472
      sex          8.13853     19.60998   5672.97963     0.17   0.6784
      v8f          1.33535      0.69987       119902     3.64   0.0572

                 Bounds on condition number: 1.222, 28.616
-------------------------------------------------------------------------------

                      Backward Elimination: Step 1

          Variable SEX Removed: R-Square = 0.1818 and C(p) = 4.1722
```

(a)

```
                            The SAS System                              37
                                   15:05 Wednesday, October 9, 2002

                           The REG Procedure
                             Model: MODEL1
                  Dependent Variable: yi 月收入:元

                      Backward Elimination: Step 1

                          Analysis of Variance

                                   Sum of          Mean
Source                  DF        Squares        Square    F Value    Pr > F

Model                    4        2716863        679216      20.67    <.0001
Error                  372       12225003         32863
Corrected Total        376       14941866


                Parameter      Standard
      Variable   Estimate         Error    Type II SS  F Value  Pr > F

      Intercept  171.42964     42.53514       533804    16.24   <.0001
      v7f         36.40275      5.07234      1692611    51.51   <.0001
      wk          -5.04739      4.63894        38905     1.18   0.2773
      edc         15.31852      7.85265       125057     3.81   0.0518
      v8f          1.34662      0.69857       122116     3.72   0.0547

                 Bounds on condition number: 1.2143, 18.164
-------------------------------------------------------------------------------

                      Backward Elimination: Step 2

          Variable wk Removed: R-Square = 0.1792 and C(p) = 3.3535
```

(b)

图 11.22　自后淘汰法的输出结果

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	2716863	679216	20.67	<.0001
Error	372	12225003	32863		
Corrected Total	376	14941866			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	171.42964	42.53514	533804	16.24	<.0001
V7F	36.40275	5.07234	1692611	51.51	<.0001
wk	-5.04739	4.63894	38905	1.18	0.2773
edc	15.31852	7.85265	125057	3.81	0.0518
v8F	1.34662	0.69857	122116	3.72	0.0547

Bounds on condition number: 1.2143, 18.164

Backward Elimination: Step 2

Variable wk Removed: R-Square = 0.1792 and C(p) = 3.3535

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	2677958	892653	27.15	<.0001
Error	373	12263908	32879		
Corrected Total	376	14941866			

(c)

The REG Procedure
Model: MODEL1
Dependent Variable: vi 月收入:元

Backward Elimination: Step 2

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	149.60289	37.51680	522816	15.90	<.0001
V7F	36.61874	5.06971	1715384	52.17	<.0001
edc	16.95516	7.70913	159042	4.84	0.0285
v8F	1.27267	0.69543	110116	3.35	0.0680

Bounds on condition number: 1.2122, 10.334

All variables left in the model are significant at the 0.1000 level.

Summary of Backward Elimination

Step	Variable Removed	Label	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	SEX	户主性别	4	0.0004	0.1818	4.1722	0.17	0.6784
2	wk	具体工作	3	0.0026	0.1792	3.3535	1.18	0.2773

(d)

图 11.22（续）

11.4 MAXR 回归法和 RSQUARE 回归法

MAXR 回归法和向前选择法有所类似，均为每步入选一个变量。所不同的是，MAXR 回归法是通过改善 R^2 值，使某个自变量驻留在模型中，详见图 11.23，每一步的 R^2 值都有所增加。然而，向前选择法一旦变量入选，不管对模型有无贡献都驻留在模型中。

RSQUARE 称为 R^2 法（有时称为所有可能法），它与多元回归法有所类似，二者均以 R^2 值的大小，来判定某变量入选后对模型贡献大小。R^2 越大，该变量入选后对模型的贡献也越大。

11.4.1 MAXR 回归法

1. MAXR 语句格式

```
PROC REG DATA=文件名 1  OUTEST=文件名 2…；          /* 详见 11.2 节 * /
MODEL 因变量=自变量表 /METHOD=MAXR SLEntry=α  (见 11.2.1 节中的图 11.11)
```

其他语句见 11.2.1 节的图 11.11。

2. 例子

例 1：

```
PROC REG DATA=d1 OUTEST=T1 CORR;
MODEL V=V1 V2/METHOD=MAXR;
PRINT ALL;
```

例 1 说明：

- 例 1 的第 1 条语句是调用 REG 过程，并计算相关矩阵。
- 第 2 条语句建立的 MODEL 为：因变量是收入（即 V），自变量是工种和工龄（即 V1 和 V2）。斜杠后面的“METHOD＝MAXR”则是本例的目的（要求用 MAXR 法进行回归分析）。
- PRINT ALL：要求显示 XPX（逆矩阵），SS1、SS2、STB、TOL、COVB 等 18 种参数，详见 11.2.2 节的注解（19）。
- 关于 MAXR 的实用例子，可进一步参阅程序 11.4 中的 MAXR 语句，以及图 11.24。

11.4.2 RSQUARE 回归法

1. RSQUARE 的语句格式

```
PROC REG DATA=文件名 2 OUTEST=T2…；(详见 11.2 节中图 11.11)
```

MODEL 因变量=自变量/METHOD=SQUARE SLEntry=α;(详见 11.2.1 节中图 11.11)

其他语句见 11.2.1 节中图 11.11。

2. 例子

例 2：

```
PROC REG DATA=d2 OUTEST=T2 CORR;
MODEL V=V1 V2/METHOD=SQUARE CP start=2 stop=3 PRINT ALL;
```

说明：

- 例 2 中的/METHOD=SQUARE CP 要求 SAS 进行所有可能法的回归分析，并显示 CP 统计量和 RSQUARE 值。
- 其余见例 1 的第 1 点至第 3 点说明。
- 实例见下述程序 11.4 的 RSQUARE 语句及其图 11.25。
- Start=2：表示被估计的自变量的最少个数。
- stop=3：表示被估计的自变量的最多个数。

11.4.3 实用程序及图例

引用 SAS 软件包 STAT SAMPLES 盘上的一个例子(文件名为 REG02EX.SAS)，以说明如何选用 MAXR 和 RSQUARE 命令进行回归分析，详见程序 11.4。

程序 11.4：

```
DATA FITNESS;
    INPUT AGE WEIGHT OXY RUNTIME RSTPULSE
          RUNPULSE MAXPULSE;
CARDS;
44 89.47 44.609 11.37 62 178 182
40 75.07 45.313 10.07 62 185 185
44 85.84 54.297 8.65 45 156 168
…
52 82.78 47.467 10.50 53 170 172
PROC REG;
    MODEL OXY=RUNTIME AGE WEIGHT RSTPULSE RUNPULSE MAXPULSE
          / METHOD=MAXR;                    /* 输出结果见图 11.24 * /
TITLE2;
     MODEL OXY=AGE WEIGHT RUNTIME RUNPULSE RSTPULSE MAXPULSE
          / METHOD=RSQUARE CP;              /* 输出结果见图 11.23~图 11.25 * /
TITLE2 'PHYSICAL FITNESS DATA: ALL MODELS';
RUN;
```

说明：程序 11.4 的 INPUT 语句中一共定义 7 个变量(见 AGE 至 MAXPULSE)。该程序用两个 MODEL 语句和两条 METHOD 命令定义了两种回归模型，每种模型

分别对应着一个输出图例。

运行程序 11.4 产生图 11.23～图 11.25 所示的结果。

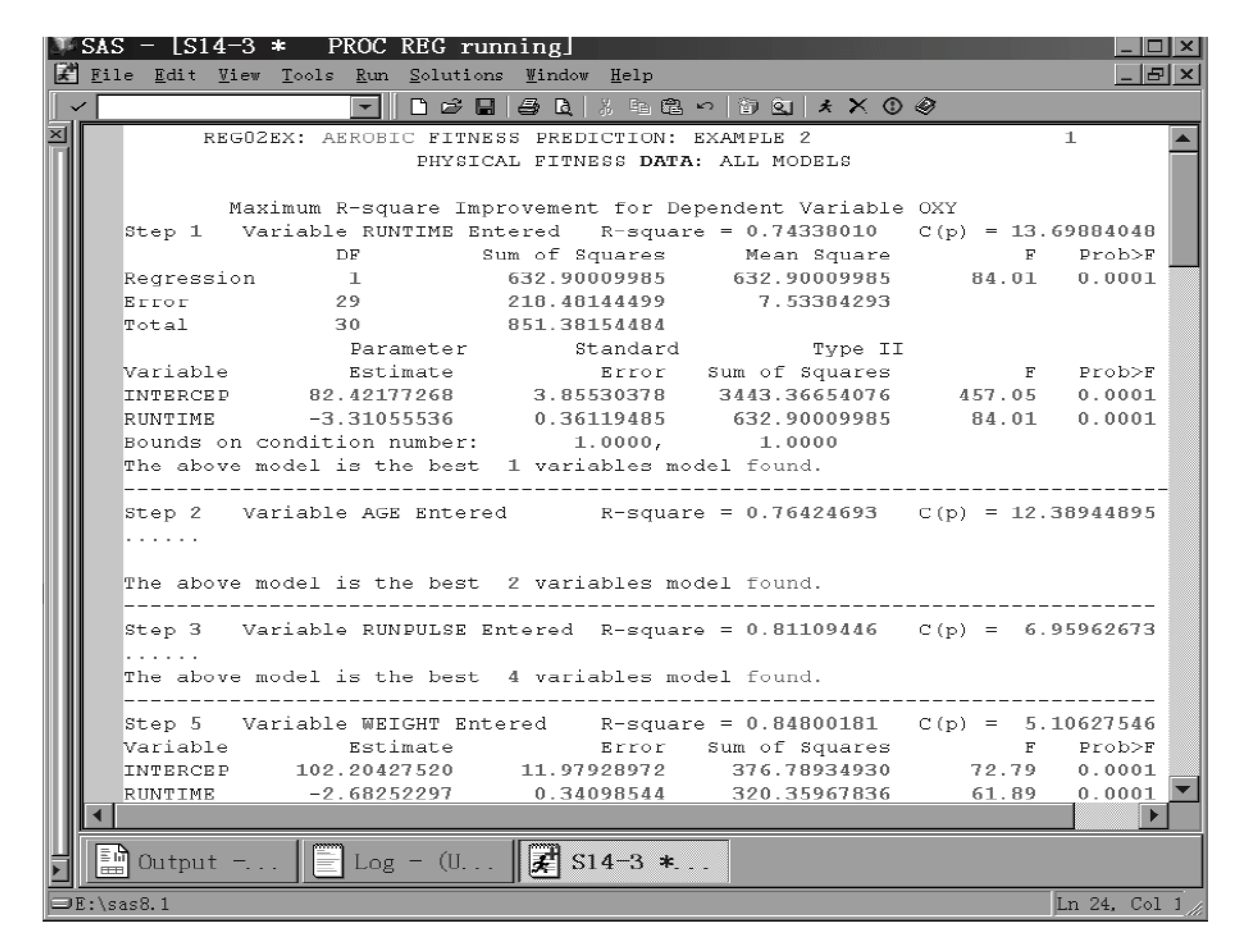


图 11.23　METHOD＝MAXR 语句产生的最大 R^2 法的输出

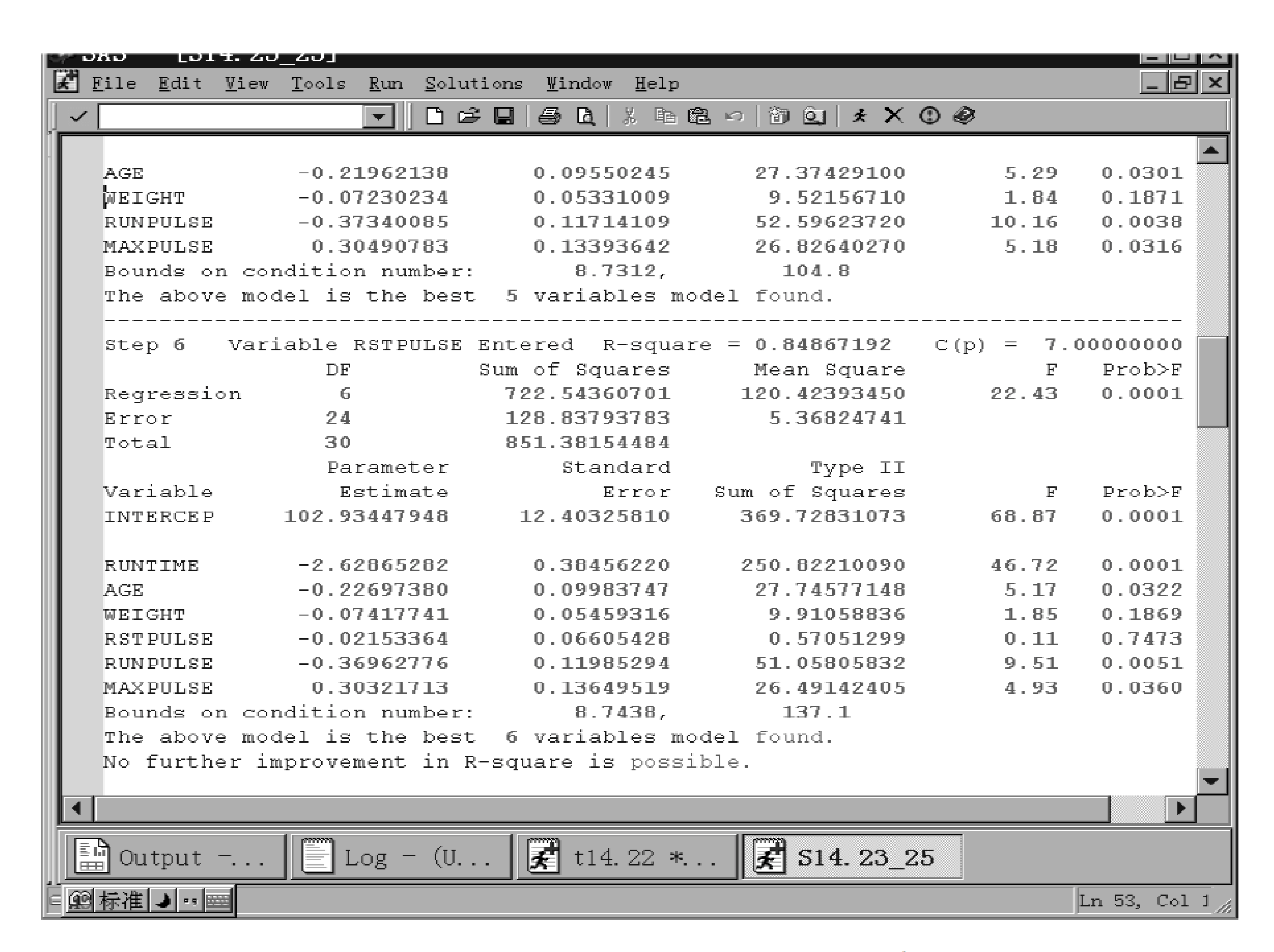


图 11.24　METHOD＝MAXR 语句产生的最大 R^2 法的输出(续)

```
N = 31   Regression Models for Dependent Variable: OXY

In   R-square    C(P)       Variables in Model

1   0.7433801   13.69884   RUNTIME
1   0.1594853   106.30211  RSTPULSE
1   0.1583834   106.47686  RUNPULSE
1   0.0927765   116.88184  AGE
1   0.0560459   122.70716  MAXPULSE
1   0.0264885   127.39485  WEIGHT
------------------------------------------------------------------------
2   0.7642469   12.389449  AGE RUNTIME
2   0.7614238   12.837184  RUNTIME RUNPULSE
2   0.7452211   15.406872  RUNTIME MAXPULSE
2   0.7449348   15.452274  WEIGHT RUNTIME
2   0.7435330   15.674598  RUNTIME RSTPULSE
2   0.3759954   73.96451   AGE RUNPULSE
2   0.3002703   85.974204  AGE RSTPULSE
2   0.2894195   87.695093  RUNPULSE MAXPULSE
2   0.2599817   92.363796  AGE MAXPULSE
2   0.2350307   96.320923  RUNPULSE RSTPULSE
2   0.1806067   104.95234  WEIGHT RSTPULSE
2   0.1740393   105.9939   RSTPULSE MAXPULSE
2   0.1668554   107.13325  WEIGHT RUNPULSE
2   0.1506353   109.70568  AGE WEIGHT
2   0.0675159   122.88807  WEIGHT MAXPULSE
------------------------------------------------------------------------
3   0.8110945   6.9596267  AGE RUNTIME RUNPULSE
3   0.8099884   7.1350367  RUNTIME RUNPULSE MAXPULSE
3   0.7817302   11.61668   AGE RUNTIME MAXPULSE
3   0.7708306   13.345306  AGE WEIGHT RUNTIME
3   0.7673494   13.897406  AGE RUNTIME RSTPULSE
3   0.7618985   14.761903  RUNTIME RUNPULSE RSTPULSE
3   0.7618290   14.772916  WEIGHT RUNTIME RUNPULSE
3   0.7461549   17.258776  WEIGHT RUNTIME MAXPULSE
3   0.7452268   17.405958  RUNTIME RSTPULSE MAXPULSE
3   0.7451114   17.424267  WEIGHT RUNTIME RSTPULSE
3   0.4666484   61.587323  AGE RUNPULSE RSTPULSE
3   0.4222735   68.625008  AGE RUNPULSE MAXPULSE
```

图 11.25　R-square 的输出示意图

3	0.4091255	70.710215	AGE WEIGHT RUNPULSE
3	0.3900068	73.742365	AGE RSTPULSE MAXPULSE
3	0.3568473	79.001324	AGE WEIGHT RSTPULSE
3	0.3537718	79.48908	RUNPULSE RSTPULSE MAXPULSE
3	0.3207793	84.721553	WEIGHT RUNPULSE MAXPULSE
3	0.2902125	89.56933	AGE WEIGHT MAXPULSE
3	0.2446512	96.79516	WEIGHT RUNPULSE RSTPULSE
…			
4	0.8116702	8.8683244	AGE RUNTIME RUNPULSE RSTPULSE
4	0.8104004	9.0697	RUNTIME RUNPULSE RSTPULSE MAXPULSE
4	0.7862243	12.903931	AGE WEIGHT RUNTIME MAXPULSE
4	0.7834321	13.346755	AGE RUNTIME RSTPULSE MAXPULSE
4	0.7750328	14.678848	AGE WEIGHT RUNTIME RSTPULSE
4	0.7622524	16.705777	WEIGHT RUNTIME RUNPULSE RSTPULSE
4	0.7461785	19.255019	WEIGHT RUNTIME RSTPULSE MAXPULSE
4	0.5033977	57.759038	AGE WEIGHT RUNPULSE RSTPULSE
4	0.5024508	57.909213	AGE RUNPULSE RSTPULSE MAXPULSE
4	0.4717197	62.783048	AGE WEIGHT RUNPULSE MAXPULSE
4	0.4256071	70.096306	AGE WEIGHT RSTPULSE MAXPULSE
4	0.3857969	76.410043	WEIGHT RUNPULSE RSTPULSE MAXPULSE

In	R-square	C(P)	Variables in Model
5	0.8480018	5.1062755	AGE WEIGHT RUNTIME RUNPULSE MAXPULSE
5	0.8370313	6.8461497	AGE RUNTIME RUNPULSE RSTPULSE MAXPULSE
5	0.8175561	9.9348366	AGE WEIGHT RUNTIME RUNPULSE RSTPULSE
5	0.8160828	10.168497	WEIGHT RUNTIME RUNPULSE RSTPULSE MAXPULSE
5	0.7887011	14.511122	AGE WEIGHT RUNTIME RSTPULSE MAXPULSE
5	0.5540659	51.723275	AGE WEIGHT RUNPULSE RSTPULSE MAXPULSE
6	0.8486719	7	AGE WEIGHT RUNTIME RUNPULSE RSTPULSE MAXPULSE

图 11.25（续）

图形分析：

1. 由于采用最大 R^2 改善法(MAXR)选取最佳的回归模型，又因为有 6 个自变量，因此图 11.24 中有 6 步回归。当每步(Step)新入选一个自变量后，R^2 都略有递增。到了最后一步，6 个自变量都进入了模型。所以回归模型可写为(取小数点后 4 位)：

$$\begin{aligned} OXY = & 102.935 - 2.6286 * Runtime - 0.2269 * AGE - 0.0742 * WEIGHT \\ & - 0.0215 * RStpulse - 0.3696 * Runpulse + 0.3032 * Maxpulse \end{aligned} \quad (11.1)$$

2. 图 11.25 是采用所有可能法(RSQUARE)进行回归分析。图中共有四栏：

(1) In：表示在方程中的变量数目。

(2) R-square：R^2 判定系数。

(3) C(P)：具有 P 个参数的回归模型之总平方误差(Total square error，简写为

TSE)，其公式为：

$$CP^* = SEEP^* \div MSEt - (n - 2P^*) \tag{11.2}$$

式(11.2)中，P^* 为 P 个自变量的回归模型的参数数目；“$SSEP^*$”为 P 个参数的回归模型的误差平方和；t 为回归模型中应包含的参数总数；MSEt 为包含 t 个参数的回归模型的均方误差；n 为样本规模即样本数目。

CP^* 参数在图 11.25 中显示为 C(P)。显见，在每一步回归分析中，C(P)值随着模型中的自变量个数的增加而增加，但 R-square 值却递减。

(4) Variables in Model：此栏见图 11.25 的第 4 栏，它表示每步回归中，模型中各个自变量的所有可能的排列组合。例如，图 11.25 中，当模型中有 5 个自变量时，该步入选的这 5 个自变量呈现着 5 种不同的先后次序，而第 5 种的排列次序为：AGE、WEIGHT、RUNPULSE、RSTPULSE、MAXPULSE，它是该步的回归结果。

到了第 6 步，也就是此例的最后一步，已进入模型的 6 个自变量，其先后顺序颇为重要，因为它们的先后次序，可作为逐步回归模型的最佳形式。

习　题　11

1. 在微型计算机 SAS 系统中 REG 过程包含哪些回归方法？
2. 请分析图 11.26 的回归结果，并写出回归预测模型。

Summary of Stepwise Selection

Step	Variable Entered	Variable Removed	Label	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	V7f		总人数	1	0.0837	0.0837	24.2831	43.11	<.0001
2	edc		文化程度	2	0.0287	0.1124	10.7788	15.25	0.0001
3	v8f		居住面积	3	0.0143	0.1268	5.0598	7.70	0.0057

09:58 Friday, June 14, 2002　3

The REG Procedure
Model: MODEL1
Dependent Variable: vi 月收入:元

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	2741025	913675	22.74	<.0001
Error	470	18884333	40179		
Corrected Total	473	21625358			

Root MSE	200.44808	R-Square	0.1268
Dependent Mean	379.81857	Adj R-Sq	0.1212
Coeff Var	52.77469		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS	Type II SS
Intercept	Intercept	1	136.45759	32.80933	4.16	<.0001	68380256	695032
V7f	总人数	1	24.09807	4.29298	5.61	<.0001	1810025	1266047
edc	文化程度	1	23.94608	7.02721	3.41	0.0007	621552	466559
v8f	居住面积	1	1.76143	0.63471	2.78	0.0057	309448	309448

图 11.26　REG 过程产生的回归结果

第 12 章 chapter 12

路径分析

路径分析是回归(Regression)分析和因素分析(Factor Analysis,也称为因子分析)的结合。路径分析是一种结构模型的分析,又称为通径分析。路径分析用以分析有时间顺序的变量中,前面的变量对随后的变量有什么影响,是通过什么"路径"(变量)影响到随后的变量等,从中可以获得"因果关系"的结构模型,这就是路径分析。因此,路径分析是一种非常重要的探索性统计方法。

在我们编著的《SPSS 实用教程》(电子工业出版社,2008 年)和《统计分析应用大全——SPSS & LISREL & SAS》(清华大学出版社,2003 年)两册教程中介绍过,有一种专门研究路径分析的单独软件叫 LISREL,它具有路径分析的强大功能。此外,SPSS 软件中也捆绑着 LISREL 模块。本章内容则是从 SAS 的角度介绍路径分析的方法。

路径分析首先是根据有关的理论知识建立一个"因果关系"的结构模型,绘制出完全模型的路径分析图(Path diagram of full Model),并找出模型中所有变量之间尽可能有的"因果关系",然后用计算线性回归系数的方法计算出路径系数。最后要剔除那些影响力甚微的变量而画出限定模型的路径分析图(Path diagram of restricted Model)。

12.1 路径分析所用的程序

本章通过对部分大学生的高考成绩及择业愿望等基本情况的调查数据,拟建一个"因果关系"的结构模型,以研究父亲职业是通过什么途径影响子女的高考成绩和择业等状况的。数据和程序文件见磁盘的 DZ9798.sas 或程序 12.1 所示。

程序 12.1:

```
OPTIONS NODATE NONUMBER;
TITLE '路径分析';
DATA PATHREG(type=corr);
INPUT _type_ $ _name_ $ vf vm score zy zo;
CARDS;
CORR  vf      1.0000    .6313   -.3548   -.2137   -.2420
CORR  vm       .6313   1.0000   -.4312   -.2239   -.1474
CORR  Score   -.3548   -.4312   1.0000    .2504    .2050
```

```
CORR   zy   -.2137   -.2239   .2504   1.0000    .1166
CORR   zo   -.2420   -.1474   .2050    .1166   1.0000
N      DF      189      189     189      189      189
;
TITLE 1 'Below is full model of path analysis';
TITLE 2 '下面是完全模型的路径分析';
PROC REG;
  MODEL vm= vf/STB;
  MODEL Score= vf vm/STB;
  MODEL zy= vf vm Score/STB;
  MODEL zo= vf vm Score zy/STB;
RUN;
TITLE1 'Below is restricted model of path analysis';
TITLE2 '下面是限定模型的路径分析';
PROC  REG;
  MODEL vm= vf/STB;
  MODEL Score= vf vm/STB;
  MODEL zy= Score/STB;
  MODEL zo= Score zy/STB;
RUN;
```

在图 12.1 中编辑程序 12.1。

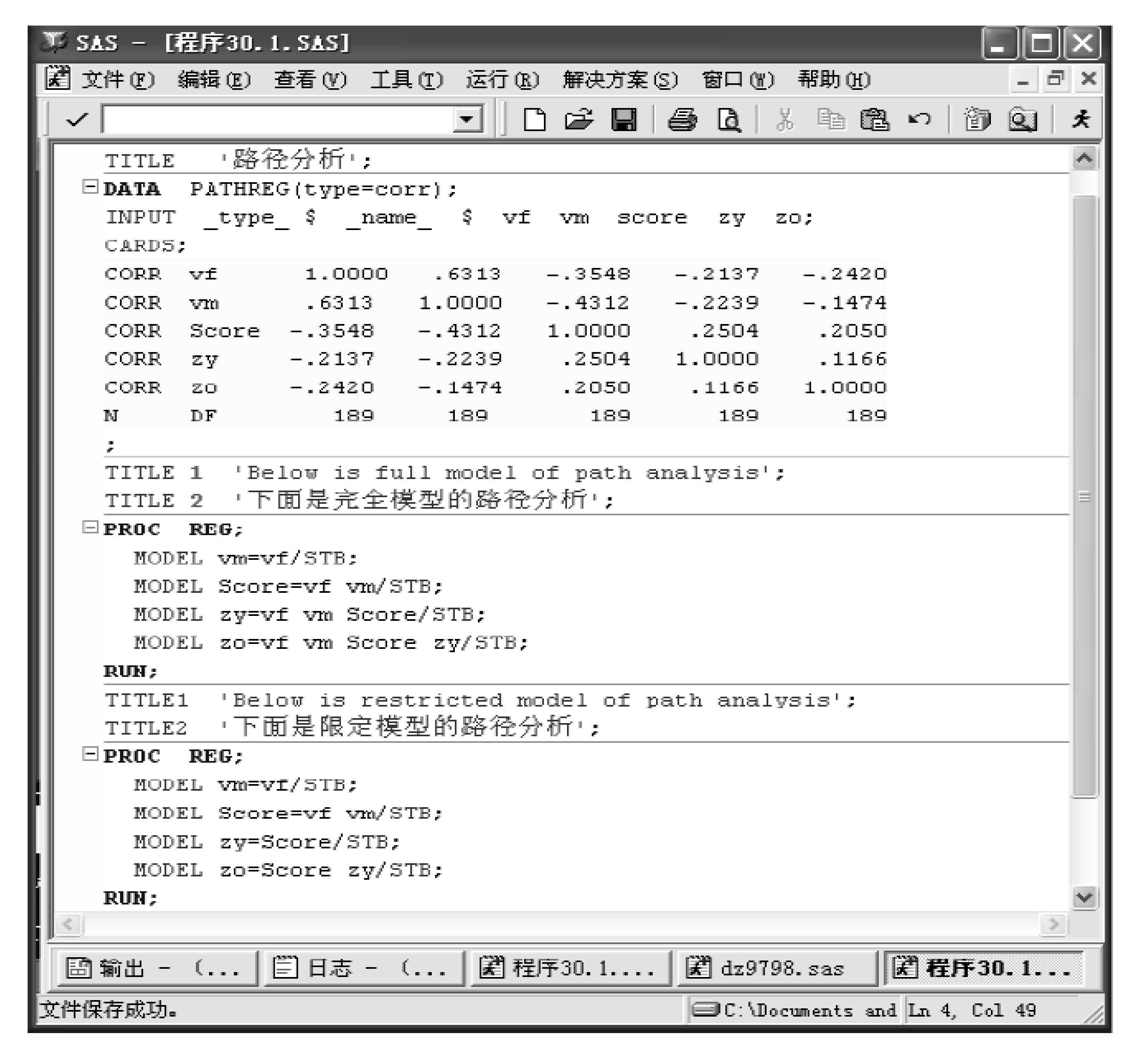


图 12.1　路径分析的程序和数据

运行程序 12.1 产生图 12.2 至图 12.5 所示的结果。但只有模块齐全且正版的 SAS 系统才具有绘制路径图的功能。

12.2 图形输出

图 12.2 是完全模型的方差分析和参数估计等统计量。由于图 12.2 的信息量大，所以分屏输出为 4 个连续图。

1. 完全模型

图 12.2 是完全模型的方差分析和参数估计。

从完全模型的方差分析和参数估计看，Score、Zy、Vm 等变量的 t 值（t Value）的概率（Pr>|t|）为 0.0776 以上，即不显著。因此本数据若用完全模型则不是理想的结构模型。

```
                    Below is full model of path analysis
                         下面是全模型的路径分析

                            The REG Procedure
                             Model: MODEL1
                        Dependent Variable: VM

                          Analysis of Variance

                                   Sum of          Mean
  Source                 DF       Squares        Square    F Value    Pr > F

  Model                   1      74.92546      74.92546     123.91    <.0001
  Error                 187     113.07454       0.60468
  Corrected Total       188     188.00000

              Root MSE              0.77761    R-Square     0.3985
              Dependent Mean              0    Adj R-Sq     0.3953
              Coeff Var                   .

                          Parameter Estimates

                     Parameter      Standard                            Standardized
Variable     DF       Estimate         Error    t Value    Pr > |t|         Estimate

Intercept     1              0       0.05656       0.00      1.0000                0
vf            1        0.63130       0.05671      11.13      <.0001          0.63130
```

(a) 最初模型拟合度检验

```
                    Below is full model of path analysis
                         下面是全模型的路径分析

                            The REG Procedure
                             Model: MODEL2
                       Dependent Variable: SCORE

                          Analysis of Variance

                                   Sum of          Mean
  Source                 DF       Squares        Square    F Value    Pr > F

  Model                   2      37.08724      18.54362      22.86    <.0001
  Error                 186     150.91276       0.81136
  Corrected Total       188     188.00000

              Root MSE              0.90075    R-Square     0.1973
              Dependent Mean              0    Adj R-Sq     0.1886
              Coeff Var                   .

                          Parameter Estimates

                     Parameter      Standard                            Standardized
Variable     DF       Estimate         Error    t Value    Pr > |t|         Estimate

Intercept     1              0       0.06552       0.00      1.0000                0
vf            1       -0.13730       0.08471      -1.62      0.1067         -0.13730
VM            1       -0.34452       0.08471      -4.07      <.0001         -0.34452
```

(b) 模型中进入了两个自变量

图 12.2 完全模型的路径分析

Below is full model of path analysis
下面是全模型的路径分析

The REG Procedure
Model: MODEL3
Dependent Variable: ZY

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	15.91193	5.30398	5.70	0.0009
Error	185	172.08807	0.93021		
Corrected Total	188	188.00000			

Root MSE	0.96447	R-Square	0.0846
Dependent Mean	0	Adj R-Sq	0.0698
Coeff Var	.		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate
Intercept	1	0	0.07015	0.00	1.0000	0
vf	1	-0.09568	0.09134	-1.05	0.2962	-0.09568
VM	1	-0.08619	0.09465	-0.91	0.3637	-0.08619
SCORE	1	0.17929	0.07851	2.28	0.0235	0.17929

(c) 模型中进入了三个自变量

Below is full model of path analysis
下面是全模型的路径分析

The REG Procedure
Model: MODEL4
Dependent Variable: ZO

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	14.84484	3.71121	3.94	0.0043
Error	184	173.15516	0.94106		
Corrected Total	188	188.00000			

Root MSE	0.97008	R-Square	0.0790
Dependent Mean	0	Adj R-Sq	0.0589
Coeff Var	.		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate
Intercept	1	0	0.07056	0.00	1.0000	0
vf	1	-0.22235	0.09214	-2.41	0.0168	-0.22235
VM	1	0.06501	0.09541	0.68	0.4965	0.06501
SCORE	1	0.14211	0.08007	1.77	0.0776	0.14211
ZY	1	0.04805	0.07395	0.65	0.5166	0.04805

(d) 最终模型

图 12.2（续）

2. 限定性模型

图 12.3 是限定性模型的方差分析和参数估计等统计量，它是分屏输出为 4 个连续图。所谓限定性模型是在完全模型的基础上剔除影响力甚微（例如系数绝对值小于 0.2）的效应项，而成为非完全模型，因此被称为限定性模型。

从图 12.3 的限定性模型的方差分析和参数估计看，Score 变量的 t 值（t Value）的概率（Pr>|t|）为 0.0120，比图 12.2 中 Score 变量的 t 值（t Value）的概率小，即变得显著了。表明限定性模型比完全模型理想。

Below is restricted model of path analysis
下面是限定模型的路径分析s

The REG Procedure
Model: MODEL1
Dependent Variable: VM

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	74.92546	74.92546	123.91	<.0001
Error	187	113.07454	0.60468		
Corrected Total	188	188.00000			

Root MSE	0.77761	R-Square	0.3985
Dependent Mean	0	Adj R-Sq	0.3953
Coeff Var	.		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate
Intercept	1	0	0.05656	0.00	1.0000	0
vF	1	0.63130	0.05671	11.13	<.0001	0.63130

(a) 最初模型拟合度检验

Below is restricted model of path analysis
下面是限定模型的路径分析s

The REG Procedure
Model: MODEL2
Dependent Variable: SCORE

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	37.08724	18.54362	22.86	<.0001
Error	186	150.91276	0.81136		
Corrected Total	188	188.00000			

Root MSE	0.90075	R-Square	0.1973
Dependent Mean	0	Adj R-Sq	0.1886
Coeff Var	.		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate
Intercept	1	0	0.06552	0.00	1.0000	0
vF	1	-0.13730	0.08471	-1.62	0.1067	-0.13730
VM	1	-0.34452	0.08471	-4.07	<.0001	-0.34452

(b) 模型2中的自变量

Below is restricted model of path analysis
下面是限定模型的路径分析s

The REG Procedure
Model: MODEL3
Dependent Variable: ZY

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	11.78763	11.78763	12.51	0.0005
Error	187	176.21237	0.94231		
Corrected Total	188	188.00000			

Root MSE	0.97073	R-Square	0.0627
Dependent Mean	0	Adj R-Sq	0.0577
Coeff Var	.		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate
Intercept	1	0	0.07061	0.00	1.0000	0
SCORE	1	0.25040	0.07080	3.54	0.0005	0.25040

(c) 模型3中的自变量

图12.3　限定性模型的路径分析

Below is restricted model of path analysis
下面是限定模型的路径分析s

The REG Procedure
Model: MODEL4
Dependent Variable: ZO

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	8.75514	4.37757	4.54	0.0119
Error	186	179.24486	0.96368		
Corrected Total	188	188.00000			

Root MSE	0.98167	R-Square	0.0466
Dependent Mean	0	Adj R-Sq	0.0363
Coeff Var	.		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate
Intercept	1	0	0.07141	0.00	1.0000	0
SCORE	1	0.18756	0.07395	2.54	0.0120	0.18756
ZY	1	0.06963	0.07395	0.94	0.3476	0.06963

(d) 最终模型

图 12.3（续）

12.3 路径图的分析方法

图 12.4、图 12.5 是示范性的两个路径图。

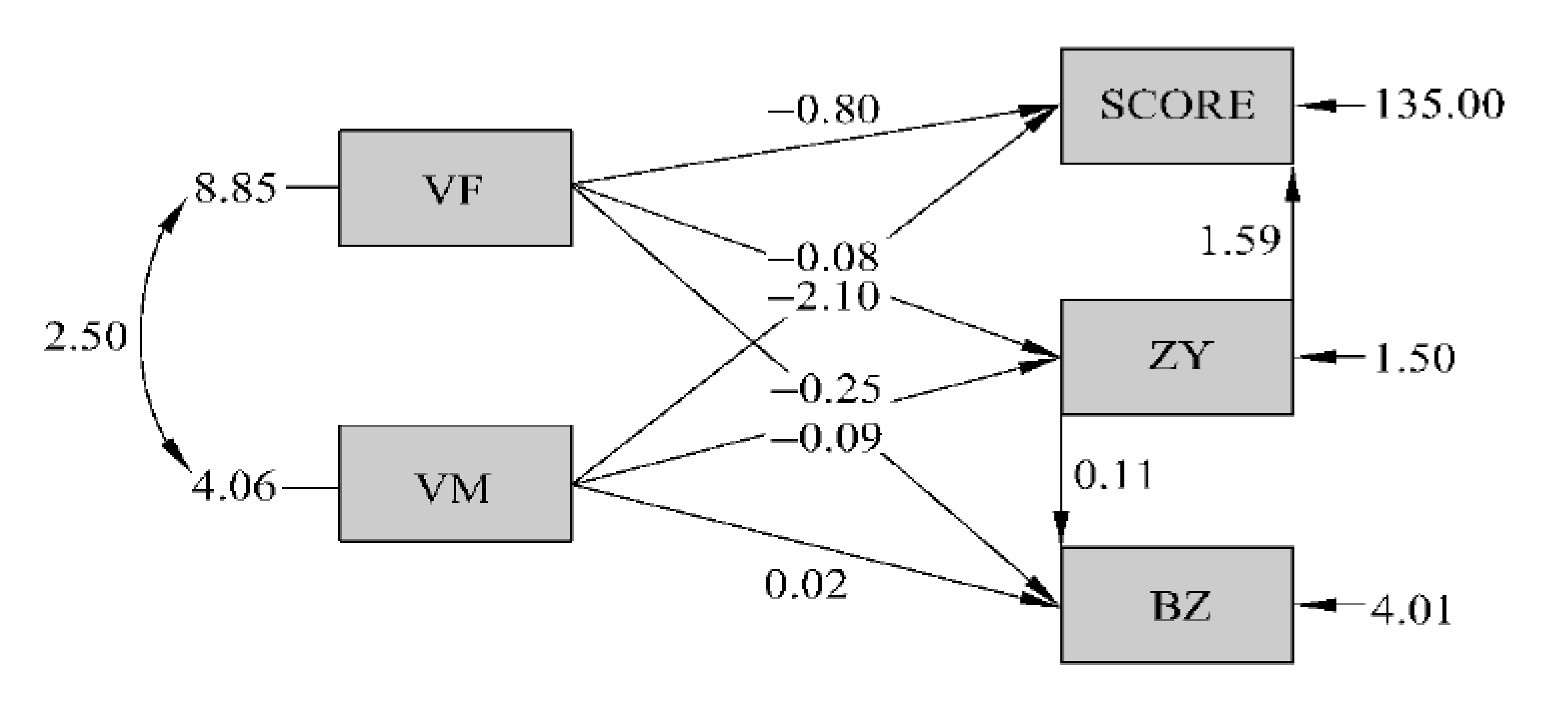


图 12.4　完全模型的路径图

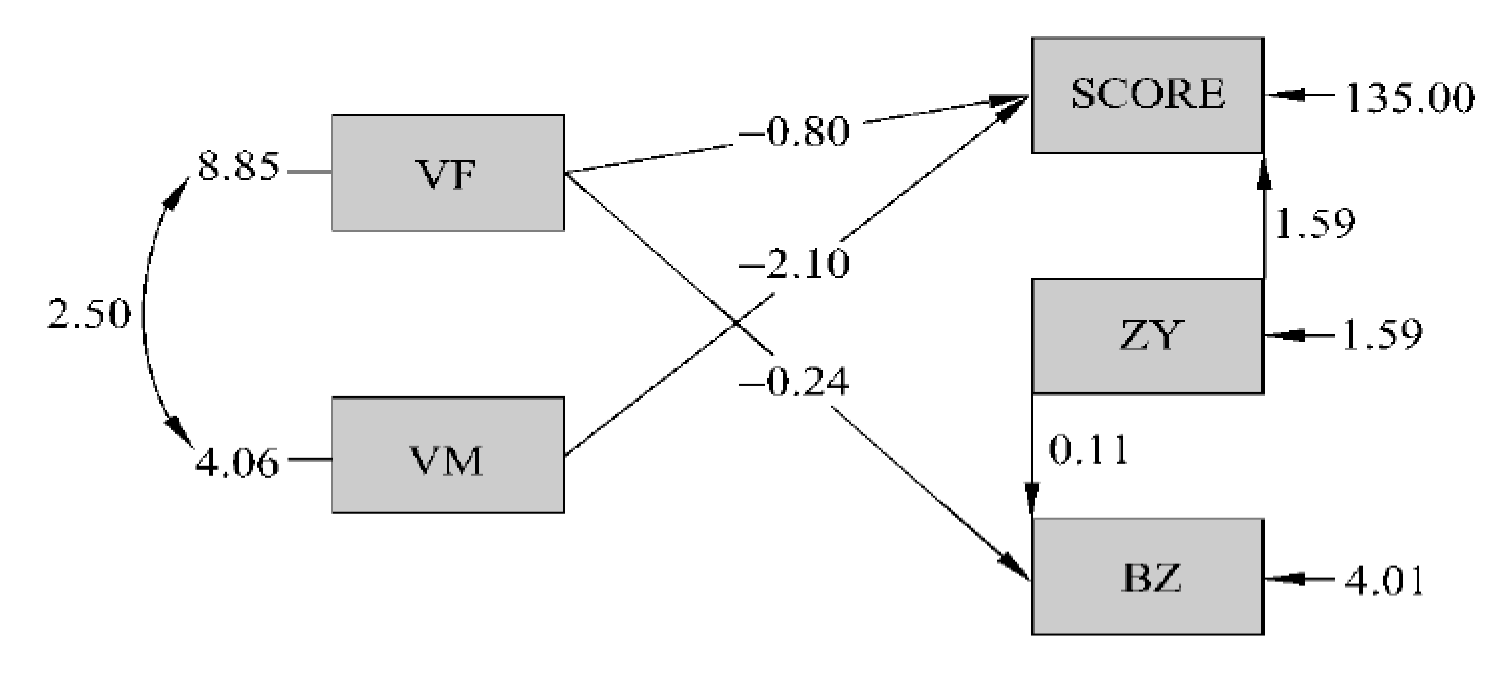


图 12.5　限定性模型的路径图

观察路径图主要是看箭头方向，它表示某变量受到来自箭头方向的变量的影响。箭头中间的数字是路径系数，即标准化的回归系数，这些系数等于图 12.2 和图 12.3 中的参数估计值。路径系数大的变量，其影响力也大。

由图 12.4 可进一步推理出母亲职业变量 Vm 的编码（如：1—工人，2—农民，3—教师，4—干部，5—医生等）越是排在后面，该生的成绩（变量 Score）则越低。而且母亲职业（变量 Vm）是通过择业（变量 zy）影响子女的择偶标准（变量 Bz）的。

由图 12.5 看出：父亲职业（变量 Vf）和学生择业（变量 Zy）两个变量则共同影响该生的成绩（Score）。

习 题 12

1. 什么是完全模型的路径图和限定性模型的路径图？

2. 请仿照第 12 章 12.1 节的程序 12.1 画出更好的“完全模型的路径图和限定性模型的路径图”。

第 13 章 chapter 13

生 存 分 析

生存分析也称为寿命分析，是对生命结束时的非追踪分析及生命进行时（未结束）的追踪分析。生命结束时由于有寿命已终结的准确时间，因此便于分析其生存周期，如一部机器的淘汰周期或一个人的寿命周期。而追踪分析由于生命仍在延续（还存在），因此无法确切地了解其生存周期（如某部机器的运行寿命为多少年，或某人手术后又活了多少年）这就需要采用本章所介绍的生存分析，即根据观察到的随机变量（人或机器）数据，进行统计与预测。在数学上，被称之为“对一个或多个非负随机变量 T 进行统计”。因而，生存分析在人类学、医学、生物学等各个学科领域有着极其重要和广泛的应用价值。

13.1 名词引论

1. 寿命函数

寿命函数其公式为：

$$S(t) = P(T > t) \tag{13.1}$$

式（13.1）中，T 是被追踪（待查）的非负随机变量 T，简称为追踪变量。

由此推论出分布函数公式为：

$$F(t) = P(T \leqslant t) \tag{13.2}$$

2. 危险率函数

另一个函数是危险率函数，即失效率或故障率函数，其公式为：

$$\lambda(t) = F(t)/S(t) \tag{13.3}$$

3. 追踪事件

追踪事件（censored cases），也称为待查或删失的个案。

4. 非追踪事件

非追踪事件（uncensored cases），也称为非待查的个案。

生存分析中的数据一般都含有追踪（待查或删失）的数据。

5. 三种追踪数据

- 左追踪数据：不知道寿命时间，只知道它大概小于 T。
- 右追踪数据：不知道寿命时间，只知道它大概大于 T。
- 区间追踪数据：不知道寿命时间 T，只知道它界于 $T_1 > T > T_2$。

6. 两种协变量

在研究生存（寿命）时，应该考虑其他因素的影响，如手术后病人的存活时间，一般是与病人的年龄、手术前的生理指标有关。这些变量被称为协变量，又称为加速变量。

协变量可以不只一个，而且有以下两类：

连续变量（如年龄、工资）和标称变量（如性别、民族）。

7. 因变量

因变量也叫响应变量。当把生存时间 T 作为因变量，把协变量 X 作为自变量时，则有：

$$Y = \mathrm{LN}(T) \tag{13.4}$$

或

$$Y = B_0 + B_1 X_1 + B_2 X_2 + \cdots + \varepsilon \tag{13.5}$$

8. 两个用于生存分析的过程命令

1）LIFEREG 过程

此过程是用参数模型去拟合“含有非追踪的生存数据”。适用于具有以下分布的数据：

指数分布、对数正态分布、对数 Logistic 分布、Gamma 分布、Weibull 分布等。

当模型中没有标称（分类）数据时，则输出数据集里的参数估计、协方差矩阵，以便进一步统计分析。

2）LIFETEST 过程

用非参数分布计算生存分布，并进行因变量与自变量相关的秩次检验。其算法采用极限乘法或寿命表法。输出的数据集里含有参数估计、协方差矩阵，以便进一步统计分析。

13.2 用 LIFEREG 进行生存分析

LIFEREG 过程是常用的生存回归过程，由参数模型去拟合可能含有追踪的数据，即含有左追踪、右追踪和区间追踪的数据。其模型由协变量的线性效应项和随机干扰项组成。随机干扰项的分布可以取以下的分布类型：

极值分布、正态分布和逻辑斯蒂克（Logistic）分布。以及通过指数转换获得的指数分布、对数正态分布、对数 Logistic 分布、Gamma 分布、Weibull 分布等。

以生存时间为因变量的模型为：

$$Y = XB + \delta\varepsilon \tag{13.6}$$

式(13.6)中,Y 是生存时间取对数值,X 是协变量(或称自变量)矩阵。B 是待求的回归系数,ε 是未知的刻度(尺度)参数。ε 被假设为标准正态分布的误差向量。这些参数可采用 Newton-Raphson 算法并通过最大似然法计算出。

13.2.1 LIFEREG过程命令

1. 过程命令的格式

```
PROC LIFEREG 选项 1;
    CLASS v1 v2;              /* v1,v2 为标称变量,即分类变量 */
    MODEL Y=变量 1/选项 2;    /* 一个 LIFEREG 过程允许多个 MODEL 语句。Y 为因变量 */
WEIGHT v3;
OUTPUT OUT=输出变量集名称 选项 3;
BY v4;
```

2. "选项 1"内容

- DATA=输入变量集名称　若省略本项,则用工作区里的数据集。
- OUTEST=数据集名称　存储参数估计值、最大似然度对数值。与 CLASS 语句互斥。
- COVOUT　在输出的数据集里存储参数估计值、最大似然度及协方差矩阵。
- NOPRINT　不在屏幕上显示统计结果。
- ORDER=FREQ|DATA|INTERNAL|FORMATTED　(仅对分类变量而言)。

其中:FREQ 要求按频数降序显示结果。DATA 要求按原始数据水平的次序显示结果。INTERNAL 要求按内部数据的格式显示结果。FORMATTED 要求按外部数据的格式显示结果。

3. MODEL 语句的格式:有以下两种。

1) 模型一

```
MODEL T(* 追踪变量 v(1 3))=v1 v2/选项 2;
```

模型一表示:T 是因变量,v 是追踪指示变量。

当 v 值为(1 3)时 T 是非追踪变量或称终检变量。当 v 值不是(1 3)时 T 为生存时间。

v1 和 v2 是协变量,但不允许它们的交互效应。协变量是分类变量时可作为主效应。

例 1:

```
MODEL T* v(1 3)=sex age;   /* v 是右追踪变量。sex 和 age 是协变量,sex 是分类变量 */
```

2) 模型二

```
MODEL (T1 T2)=v1 v2/选项 2;
```

模型二表示：T1 是追踪变量的下界，若缺失 T1，则用上界 T2 作为左追踪值。T2 是追踪变量的上界，若缺失 T2，则用下界 T1 作为右追踪值。

例 2：

```
MODEL (T1 T2)= sex age;     /* T1 和 T2 是因变量的范围，形成区间追踪。sex 和 age
                               是协变量 */
```

例 3：

```
MODEL (T1 T2)=;             /* T1 和 T2 是区间追踪。只有默认的截距项而没有
                               协变量 sex age */
```

3）模型三

```
MODEL events/trials= v1 v2/选项 2;
```

模型三表示：events 为成功的次数，trials 为试验次数。

例 4：

```
MODEL gz/T= sex age;        /* 假设 gz 为成功次数，T 为试验次数 */
```

4. MODEL 语句中的“选项 2”内容

MODEL 中的“选项 2”有以下 3 类：

1）模型选择

- D 值的设定：

```
D= Weibull          /* D 是 Distribution 的缩写。此处采用 Weibull 分布 */
D= Exponential      /* 采用指数分布 */
D= Llogistic        /* 采用对数逻辑斯蒂克分布 */
D= Logistic         /* 采用逻辑斯蒂克分布，类同于用 Nolog 时的对数逻辑斯蒂克分布 */
D= Gamma            /* 采用伽马分布 */
D= Normal           /* 采用正态分布，类同于用 Nolog 时的对数正态分布 */
D= Lnormal          /* 采用对数正态分布 */
```

- NOLOG　不对因变量进行对数转换。
- INTERCPT＝指定值　　要求截距项改为“指定值”。
- NOINT　若无初始的截距项，则截距项置 0 值。鉴于常对因变量做对数转换，截距项通常为没有变换的因变量的常量。
- INILIAL＝初始值　给协变量回归系数设置初始值。收敛困难时，很有用。

以下 4 项用于参数说明。

- SCALE＝值　用该“值”作为参数的初始值。对于指数模型，它同固定的 SCALE ＝1 Weibull 模型。
- NOSCALE　要求刻度参数保持固定。若不指定 SCALE＝值，则刻度默认为 1。若生存变量为对数转换，则刻度参数的作用是原始响应的幂转换。
- SHAPEI＝值　形状分布（参数）的初始值。如果规定的分布与此参数无关，则此

项不起作用。

• NOSHAPE1　第一个形状分布(参数)必须固定。

2）模型拟合度的选项

• CONVERGE＝值　收敛值。默认(即预置值)为 0.001。
• MAXIT＝值　允许迭代的最大次数,默认为 50 次。
• SINGULAR＝n　检验枢纽值,它至少是原始对角值的 n 倍。默认为 1E－12。

3）输出的选项

• CORRB　要求显示参数估计的相关矩阵。
• COVB　用逆矩阵作为参数的协方差矩阵。
• ITPRINT　显示迭代过程及梯度和第二阶微商矩阵(hessian)的最终估计。

5．输出语句：OUTPUT DATA＝名 选项 3

其中选项 3 的内容如下：

• CENSRED＝变量 v　变量 v 为追踪的指示变量名。若个案中含有追踪则 v 值取 1,否则 v＝0。
• CDF＝v1　将累积分布函数的估计值存入 v 变量中。
• CONTROL＝v2　这里的 v2 为数据集里的变量,用来控制分位数估计值。当 v2＝1 时则对所在的个案计算“QUANTILE＝值 1 值 2”范围的估计值。否则内部计算估计值。若省略 CONTROL 项,则计算全部个案的所有分位数。若有第三种模型“MODEL C/n＝”项,则不能使用“CONTROL＝ v2”项。
• OUT＝D1　指定数据集名称为 D1。若省略此项,则用 DATA　n 作为数据集名称。
• P(或 Predict)＝v3　将分位数估计值(或概率值)存入 v3 变量中。对于第三种模型“MODEL C/n＝”,则计算 $1-F(-X'\beta)$ 值(即概率值)。
• Q(Quantiles)＝值 1 值 2　指定需要计算的分位数。分位数必须在 0～1 之间(不含 0 和 1)。默认为 Q＝0.5。对于第三种模型“MODEL C/n＝”,则不能指定“Q＝值 1 值 2”项。
• STD_ERR_＝v4(或用 STD＝v4)　将分位数估计值或 $X'\beta$ 的标准误差存入 v4 变量中。对于“MODEL C/n＝”模型,则计算 $X'\beta$ 的标准误差。这些估计可用来计算分位数的置信区间。
• XBATE＝v5　将 $X'\beta$ 的计算值存入 v5 中。X 为协方差向量,B 为参数估计的向量。

6．条件限制

1）缺失值的处理

若一个个案的因变量(或某个自变量,或 sensor 追踪变量)含有缺失值,则回归分析时删除该个案。如自变量不含缺失值,则继续计算其预测值。

2）不允许 MODEL 语句指定交互作用

在 LIFEREG 过程中指定“MODEL Y＝v1 v2”是正确的;但若指定“MODEL Y＝v1

v2 v1 * v2"则是错误的。

13.2.2 LIFEREG过程的应用实例

例 5：引用英文 SAS 中的发动机失效例子。

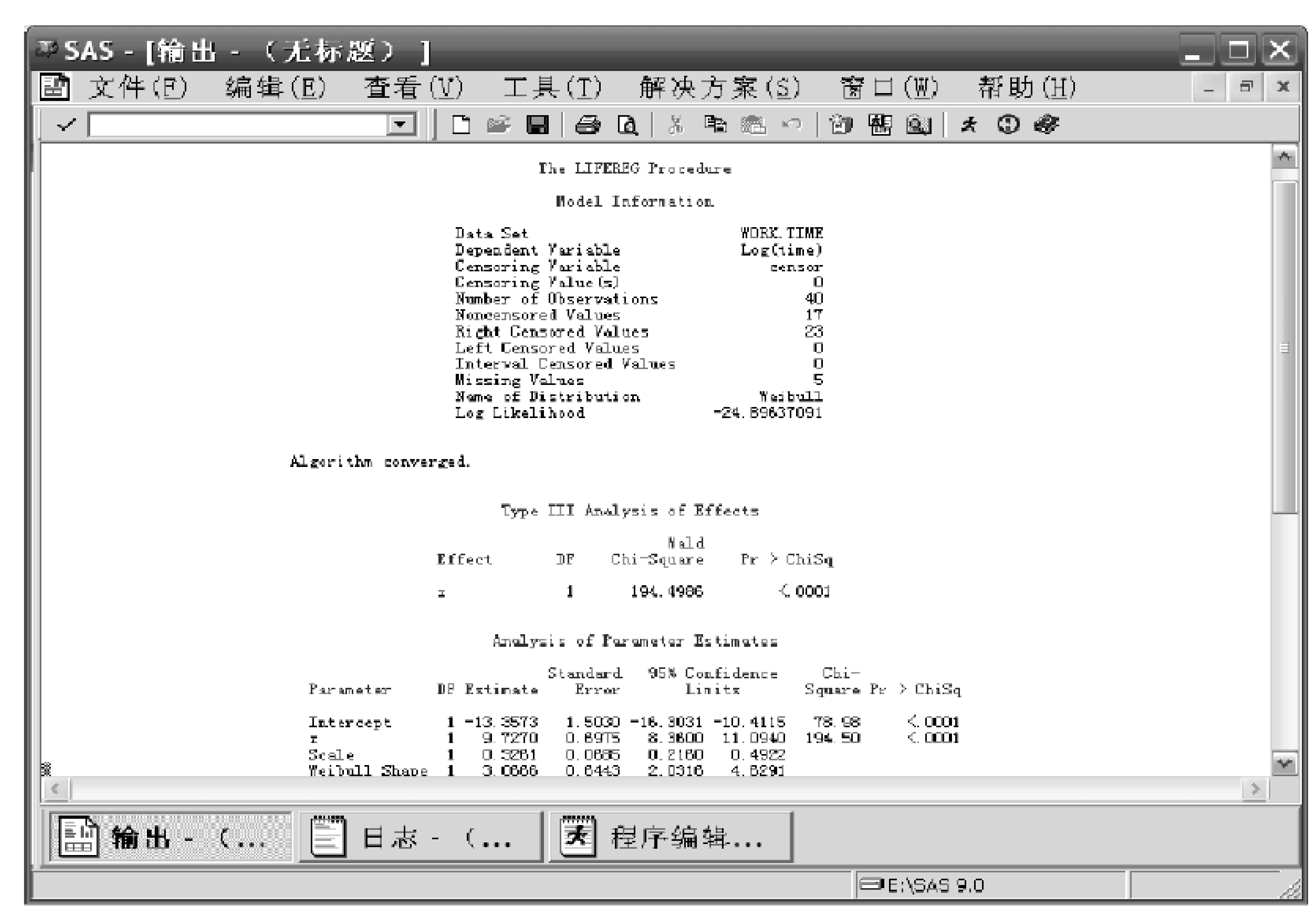
这是一个发动机因为运行发热(温度)使用了一段时间便失效的生命分析实例。这是美国 Kalbfleisch 与 Prentice(1980.5)提供的,并用 Weibull 模型及对数正态模型来拟合数据的。数据见程序 13.1。

程序 13.1：发动机数据。

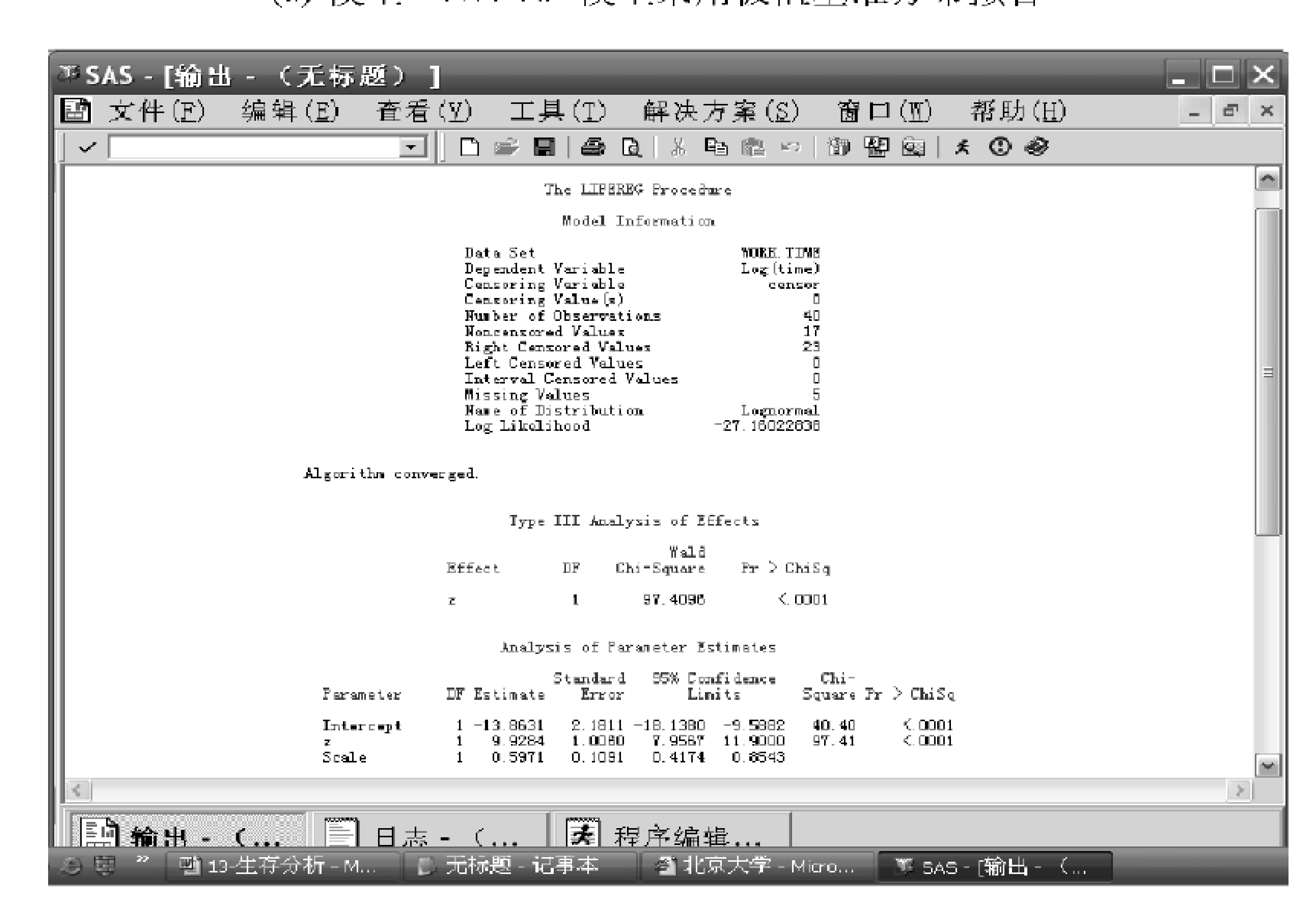
```
DATA  TIME;
    * 以下的 time 是发动机寿命。censor 是追踪的指示变量,temp 是工作温度;
INPUT time censor temp@@;            /* censor=1 时,time 是寿命时间。censor=0 时,time 是右追踪
                                        变量 */
z=1000/(273.2+temp);
CONTROL=(_N_>40);                    /* N 为内部控制变量,代表个案号的序号。当_N_等于 41~45
                                        时,变量 CONTROL 值为 1,并输出本例数据最后 5 个个案。
                                        否则 CONTROL 为 0 */
CARDS;
8064  0  150  8064  0  150  8064  0  150  8064  0  150  8064  0  150
8064  0  150  8064  0  150  8064  0  150  8064  0  150  8064  0  150
1764  1  170  2772  1  170  3442  1  170  3542  1  170  3780  1  170
4860  1  170  5196  1  170  5448  0  170  5448  0  170  5448  0  170
 408  1  190   408  1  190  1334  1  190  1334  1  190  1440  1  190
1680  0  190  1680  0  190  1680  0  190  1680  0  190  1680  0  190
0408  1  220  0408  1  220  0504  1  220  0504  1  220  0504  1  220
0528  0  220  0528  0  220  0528  0  220  0528  0  220  0528  0  220
   .  0  130     .  0  150     .  0  170     .  0  190     .  0  220
PROC  LIFEREG OUTEST=MODELS;
  A:MODEL TIME*CENSOR(0)=z;          /* censor=0 时,time 是右追踪 */
  * 下面对因变量进行对数转换。Weibull 模型采用极值基准分布拟合。对数正态模型用正态基
  准分布拟合;
  B: MODEL TIME*CENSOR(0)=z/DIST= LNORMAL;
    * 下面语句要求计算三个百分位数的标准差、预测值,只对 CONTROL= 1 的个案计算分位数及
    其他值;
OUTPUT OUT=OUT1 QUANTILES=.1 .5 .9 STD_ERR= STD
    P=PREDTIME CONTROL=CONTROL;
PROC PRINT DATA=MODELS;
ID _MODEL_;                          /* 特殊变量_MODELS_= A时表示个案来自模型 A* /
DATA time;                           /* 响应的 95% 置信区间 */
    SET OUT1;                        /* 从 OUTPUT 语句存储到的数据集 OUT1 中,复制标准差的估计
                                        值 */
LTIME=LOG(PREDTIME);                 /* 将预测值取对数 */
```

```
  stde= STD/PREDTIME;
UPPER= EXP(LTIME+ 1.6 * stde);          /* 把置信区间还原为原来的刻度 * /
LOWER= EXP(LTIME- 1.6 * stde);
PROC PRINT;
    ID temp;                             /* 用 temp 识别个案 * /
TITLE '分位数估计和置信区间估计';
RUN;
```

运行程序 13.1 产生图 13.1 所示的结果。

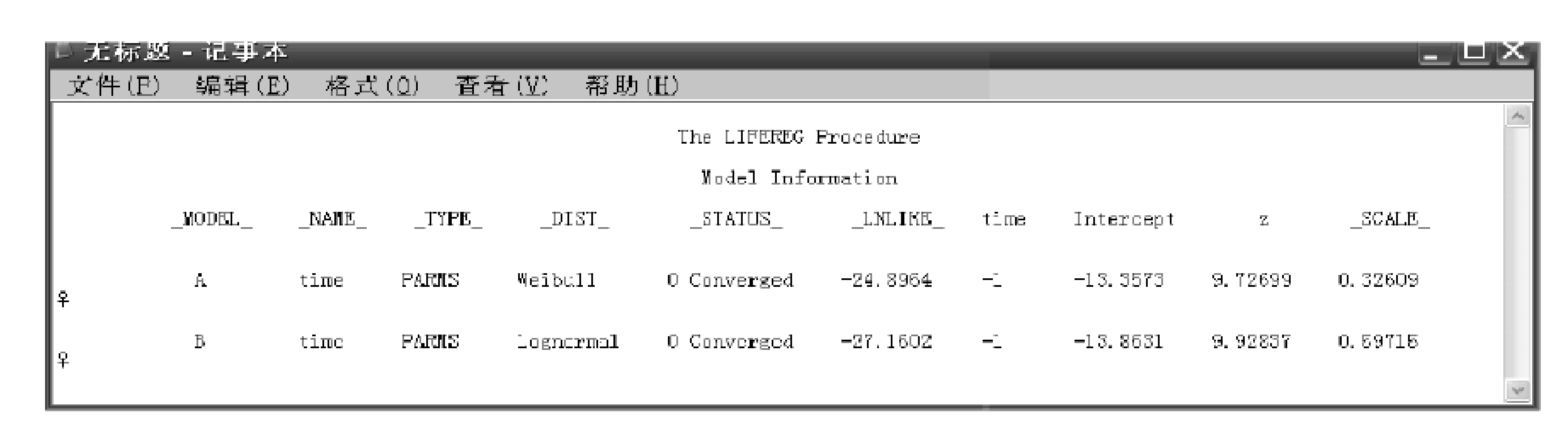


(a) 模型 A:Weibull模型采用极值基准分布拟合



(b) 模型 B:对数正态模型用正态基准分布拟合

图 13.1　发动机寿命分析


无标题 - 记事本

The LIFEREG Procedure

Model Information

MODEL	_NAME_	_TYPE_	_DIST_	_STATUS_	_LNLIKE_	time	Intercept	z	_SCALE_
A	time	PARMS	Weibull	0 Converged	-24.8964	-1	-13.3573	9.72699	0.32609
B	time	PARMS	Lognormal	0 Converged	-27.1602	-1	-13.8631	9.92837	0.59715

(c) 数据集 MODELS的信息

SAS - [输出 - (无标题)]

分位数估计和置信区间估计　　15:23 Friday, April 26, 2002

temp	time	censor	z	CONTROL	_PROB_	PREDTIME	STD	LTIME	stde	UPPER	LOWER
130	.	0	2.48016	1	0.1	21928.39	6959.26	9.9955	0.31736	36436.20	13197.16
130	.	0	2.48016	1	0.5	47137.00	16134.10	10.7608	0.34228	81508.56	27259.68
130	.	0	2.48016	1	0.9	101325.12	42107.14	11.5261	0.41556	197006.70	52113.86
150	.	0	2.36295	1	0.1	6848.85	1480.83	8.8318	0.21622	9679.65	4845.91
150	.	0	2.36295	1	0.5	14722.20	3489.48	9.5971	0.23702	21511.57	10075.66
150	.	0	2.36295	1	0.9	31646.67	10240.49	10.3624	0.32359	53110.47	18857.14
170	.	0	2.25632	1	0.1	2375.97	350.84	7.7732	0.14766	3009.17	1876.01
170	.	0	2.25632	1	0.5	5107.36	805.26	8.5384	0.15767	6572.85	3968.61
170	.	0	2.25632	1	0.9	10978.72	2842.07	9.3037	0.25887	16612.43	7255.54
190	.	0	2.15889	1	0.1	903.17	124.92	6.8059	0.13831	1126.87	723.87
190	.	0	2.15889	1	0.5	1941.44	248.32	7.5712	0.12791	2382.33	1582.15
190	.	0	2.15889	1	0.9	4173.30	957.96	8.3365	0.22954	6025.37	2890.52
220	.	0	2.02758	1	0.1	245.21	52.55	5.5021	0.21429	345.50	174.03
220	.	0	2.02758	1	0.5	527.10	99.03	6.2674	0.18789	711.95	390.25
220	.	0	2.02758	1	0.9	1133.05	286.15	7.0327	0.25255	1697.23	756.42

(d) 各个个案的预期寿命结果

图 13.1（续）

结果分析：

在输出结果中除了 OUT 语句中的变量外，还有 DATA 步所建立的新变量 LTIME、STDE、UPPER、LOWER。

（1）回归方程可靠性分析

从图 13.1 的(a)和(b)两小图看，“Pr > ChiSq”为 0.0001 以下，这两个概率值都分别小于 α 值 0.05，说明两个模型都拟合数据。

由此可以根据图 13.1(a)写出模型 A 的寿命回归方程，根据图 13.1(b)写出模型 B 的寿命回归方程。

（2）z 系数

z 系数是将摄氏温度转换为绝对温度后取倒数。从图 13.1(c)看，模型 A 的 z 值为 9.72699，模型 B 的 z 值为 9.92837。

（3）置信区间

图 13.1(d)是模型 B 的置信区间，它给出了各种温度下，10%、50%、90%分位数(见 PROB 栏)的预测寿命，以及 90%置信区间的上下限。从图 13.1(d)看，温度为 220 度时，90%分位数的预测寿命为 1133.05。

例 6：引自英文 SAS 中肺癌患者的生存数据，T 为寿命时间。变量见表 13.1，数据见程序 13.2(a)，试分析其寿命。

根据表 13.1 的变量和题意，编辑出的程序 13.2(a)如下。

程序 13.2(a)：

表 13.1　肺癌患者的生存变量

变　量　名	变 量 取 值
CELL(细胞类型)	1＝鱼鳞状　2＝小型　3＝大型　4＝腺状
THERAPY(疗法)	1＝标准(standard)　2＝试验(test)
PRIOR(曾治疗否)	1＝治疗过(yes)　2＝未治疗(no)
AGE(年龄:岁)	__岁
DIAGTIME(从诊断到治疗的年数)	__年
KPS(手术前的综合体质)	__评分

```
DATA valung;
LABEL T= '追踪时间或非追踪时间' kps= '手术前的综合素质评分'
      diagtime= '从诊断到手术治疗的时间'
      age= '年龄' prior= '事先是否有治疗' cell= '细胞类型'
      therapy= '疗法';
INPUT therapy cell t kps diagtime age prior $  @@ ;
    CENSOR= (t< 0);
    t= ABS(t);
CARDS;
1  1   072  60   7  69 n 1 1   411  70  05  64 y 1 1   228  60   3  38 n
1  1   126  60   9  63 y 1 1   118  70  11  65 y 1 1    10  20   5  49 n
1  1   082  40  10  69 y 1 1   110  80  29  68 n 1 1   314  50  18  43 y
1  1  -100  70  06  70 n 1 1   042  60  04  81 n 1 1   008  40  58  63 y
1  1   144  30   4  63 n 1 1   -25  80   9  52 y 1 1    11  70  11  48 y
1  2    30  60   3  61 n 1 2   384  60   9  42 n 1 2    04  40  02  35 n
1  2    54  80   4  63 y 1 2    13  60   4  56 n 1 2  -123  40  03  55 n
1  2   -97  60   5  67 n 1 2   153  60  14  63 y 1 2    59  30   2  65 n
1  2   117  80   3  46 n 1 2   016  30  04  53 y 1 2   151  50  12  69 n
1  2    22  60   4  68 n 1 2    56  80  12  43 y 1 2    21  40   2  55 y
1  2    18  20  15  42 n 1 2   139  80  02  64 n 1 2    20  30   5  65 n
1  2    31  75   3  65 n 1 2   052  70  02  55 n 1 2   287  60  25  66 y
1  2    18  30   4  60 n 1 2    51  60   1  67 n 1 2   122  80  28  53 n
1  2    27  60   8  62 n 1 2    54  70   1  67 n 1 2   007  50   7  72 n
1  2    63  50  11  48 n 1 2   392  40  04  68 n 1 2    10  40  23  67 y
1  4    08  20  19  61 y 1 4    92  70  10  60 n 1 4    35  40   6  62 n
1  4   117  80  02  38 n 1 4   132  80   5  50 n 1 4    12  50   4  63 y
1  4   162  80   5  64 n 1 4   003  30  03  43 n 1 4    95  80   4  34 n
1  3   177  50  16  66 y 1 3   162  80   5  62 n 1 3   216  50  15  52 n
1  3   553  70   2  47 n 1 3   278  60  12  63 n 1 3   012  40  12  68 y
1  3   260  80   5  45 n 1 3   200  80  12  41 y 1 3   156  70   2  66 n
1  3  -182  90   2  62 n 1 3   143  90   8  60 n 1 3   105  80  11  66 n
1  3   103  80   5  38 n 1 3   250  70   8  53 y 1 3   100  60  13  37 y
2  1   999  90  12  54 y 2 1   112  80   6  60 n 2 1   -87  80   3  48 n
2  1  -231  50   8  52 y 2 1   242  50   1  70 n 2 1   991  70   7  50 y
2  1   111  70   3  62 n 2 1   001  20  21  65 y 2 1   587  60   3  58 n
```

```
2  1  389  90   2  62 n 2 1   033 30 06 64 n 2 1   25 20 36 63 n
2  1  357  70  13  58 n 2 1   467 90  2 64 n 2 1  201 80 28 52 y
2  1  001  50   7  35 n 2 1    30 70 11 63 n 2 1  044 60 13 70 y
2  1  283  90   2  51 n 2 1    15 50 13 40 y
2  2   25  30   2  69 n 2 2  -103 70 22 36 y 2 2   21 20 04 71 n
2  2   13  30   2  62 n 2 2   087 60 02 60 n 2 2   02 40 36 44 y
2  2   20  30   9  54 y 2 2   007 20 11 66 n 2 2   24 60  8 49 n
2  2   99  70   3  72 n 2 2   008 80 02 68 n 2 2   99 85  4 62 n
2  2   61  70   2  71 n 2 2   025 70 02 70 n 2 2   95 70  1 61 n
2  2   80  50  17  71 n 2 2   051 30 87 59 y 2 2   29 40  8 67 n
2  4   24  40  02  60 n 2 4   018 40 05 69 y 2 4  -83 99  3 57 n
2  4   31  80  03  39 n 2 4   051 60 05 62 n 2 4   90 60 22 50 y
2  4   52  60  03  43 n 2 4   073 60 03 70 n 2 4   08 50 05 66 n
2  4   36  70  08  61 n 2 4   048 10 04 81 n 2 4   07 40 04 58 n
2  4  140  70  03  63 n 2 4   186 90 03 60 n 2 4   84 80  4 62 n
2  4  019  50  10  42 n 2 4    45 40 03 69 n 2 4   80 40 04 63 n
2  3  052  60  04  45 n 2 3   164 70 15 68 y 2 3   19 30 04 39 y
2  3  053  60  12  66 n 2 3   015 30 05 63 n 2 3   43 60 11 49 y
2  3  340  80  10  64 y 2 3   133 75 01 65 n 2 3  111 60 05 64 n
2  3  231  70  18  67 y 2 3   378 80 04 65 n 2 3  049 30 03 37 n
;
PROC FORMAT;
  VALUE cellf 1='鱼鳞状'  2='小型'  3='大型'  4='腺状';
  FORMAT cell  cellf.;
PROC LIFEREG;
    CLASS therapy cell  prior;
MODEL t * CENSOR(1)=therapy cell  prior kps age diagtime/D=WEIBULL;
OUTPUT OUT= OUT2 P=PRED;
RUN;
```

程序 13.2(a)说明：

CENSOR=(t<0)：指定 t<0 的个案为追踪(失访,还活着)数据。

t=ABS(t)：将因变量值全部转换为正值。

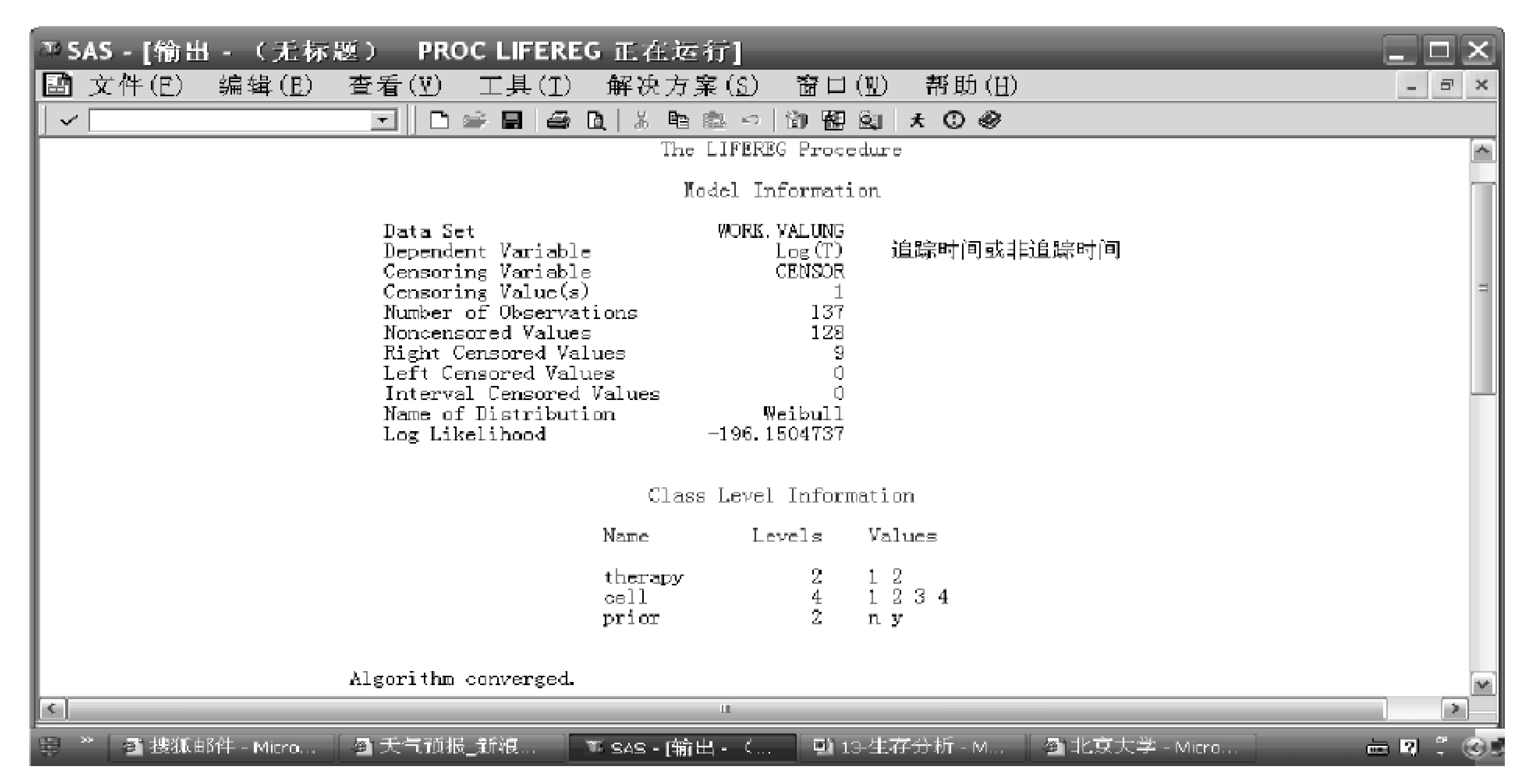
CLASS：指定 cell 变量等为分类变量。

MODEL t * CENSOR(1)语句：指定 t 为因变量(寿命时间)、CENSOR(1)为追踪变量、1 为追踪指示值。即当 t 为负值时为追踪值(CENSOR=1)。当 t 为正值时，为非追踪值(CENSOR=0)。

运行程序 13.2(a)产生图 13.2 所示的结果。

对图 13.2(b)结果的分析：

LIFEREG 输出结果的核心部分要看每个变量的“Pr > ChiSq”值。从图 13.2(b)看，截距 Intercept 的该 Pr 值 0.0071 小于 α 值 0.05，很显著。cell(细胞类型)变量和 kps(手术前的综合体质)变量的 Pr > ChiSq 值都小于 0.0001，因此也都小于 α 值 0.05，非常显著。而其余变量的“Pr > ChiSq”值都不显著。所以模型中只需要保留 cell、kps 和


SAS - [输出 - （无标题） PROC LIFEREG 正在运行]

The LIFEREG Procedure

Model Information

Data Set	WORK.VALUNG	
Dependent Variable	Log(T)	追踪时间或非追踪时间
Censoring Variable	CENSOR	
Censoring Value(s)	1	
Number of Observations	137	
Noncensored Values	128	
Right Censored Values	9	
Left Censored Values	0	
Interval Censored Values	0	
Name of Distribution	Weibull	
Log Likelihood	-196.1504737	

Class Level Information

Name	Levels	Values
therapy	2	1 2
cell	4	1 2 3 4
prior	2	n y

Algorithm converged.

(a) Weibull分布的概况

SAS - [输出 - （无标题） PROC LIFEREG 正在运行]

Type III Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
therapy	1	1.5572	0.2121
cell	3	20.3261	0.0001
prior	1	0.0189	0.8908
kps	1	38.9454	<.0001
age	1	0.5249	0.4688
diagtime	1	0.0453	0.8314

Analysis of Parameter Estimates

Parameter		DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept		1	1.9325	0.7174	0.5264	3.3385	7.26	0.0071
therapy	1	1	0.2325	0.1863	-0.1327	0.5976	1.56	0.2121
therapy	2	0	0.0000	.	.	.	.	.
cell	1	1	1.1135	0.2621	0.5998	1.6272	18.05	<.0001
cell	2	1	0.3050	0.2514	-0.1878	0.7978	1.47	0.2252
cell	3	1	0.7230	0.2741	0.1859	1.2602	6.96	0.0083
cell	4	0	0.0000	.	.	.	.	.
prior	n	1	-0.0302	0.2199	-0.4612	0.4008	0.02	0.8908
prior	y	0	0.0000	.	.	.	.	.
kps		1	0.0299	0.0048	0.0205	0.0392	38.95	<.0001
age		1	0.0062	0.0086	-0.0106	0.0230	0.52	0.4688
diagtime		1	-0.0018	0.0084	-0.0183	0.0147	0.05	0.8314
Scale		1	0.9277	0.0616	0.8146	1.0566		
Weibull Shape		1	1.0779	0.0715	0.9464	1.2276		

(b) 基准分布为Weibull分布的模型检验

图 13.2 拟合基准分布为 Weibull 分布的输出结果

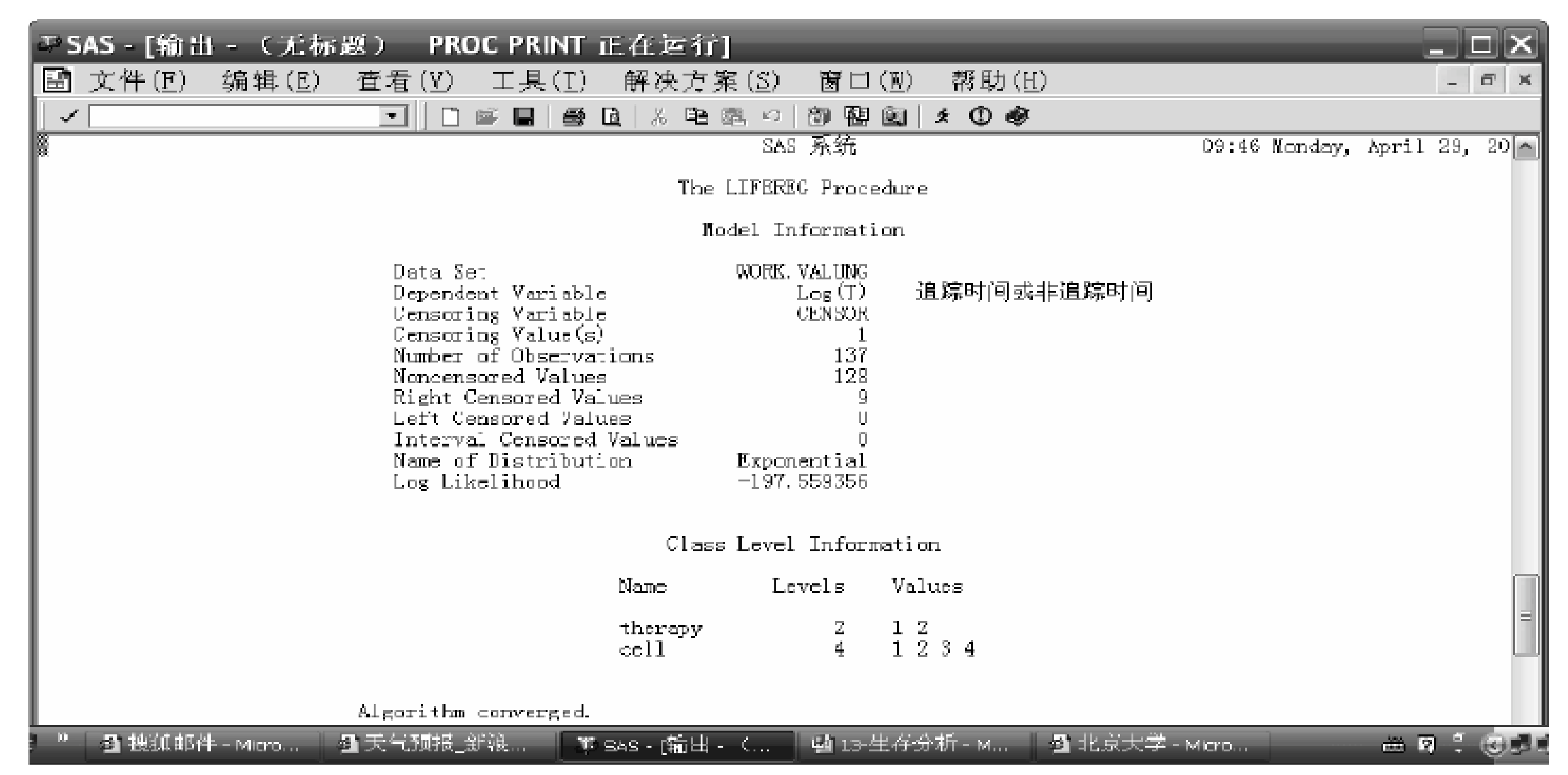
Intercept 三个变量。为此，对程序 13.2(a)最后几条语句(特别是“MODEL t * CENSOR(1)=therapy cell prior kps age diagtime/D=WEIBULL;”)换成如下语句(见程序 13.2(b))。

程序 13.2(b)：

(注：前面的语句与程序 13.2(a)相同)

```
PROC LIFEREG;
    CLASS therapy cell;
MODEL t * CENSOR(1)=cell kps/EXPONENTIAL;
OUTPUT OUT=OUT2 P=PRED;
PROC PRINT;
```

然后重新运行程序 13.2(b)产生简要的输出结果如图 13.3、图 13.4 所示。



```
SAS - [输出 - (无标题)   PROC PRINT 正在运行]
文件(F) 编辑(E) 查看(V) 工具(T) 解决方案(S) 窗口(W) 帮助(H)
                                SAS 系统                    09:46 Monday, April 29, 20

                          The LIFEREG Procedure

                           Model Information

          Data Set                      WORK.VALUNG
          Dependent Variable                  Log(T)    追踪时间或非追踪时间
          Censoring Variable                  CENSOR
          Censoring Value(s)                       1
          Number of Observations                 137
          Noncensored Values                     128
          Right Censored Values                    9
          Left Censored Values                     0
          Interval Censored Values                 0
          Name of Distribution           Exponential
          Log Likelihood                 -197.559356

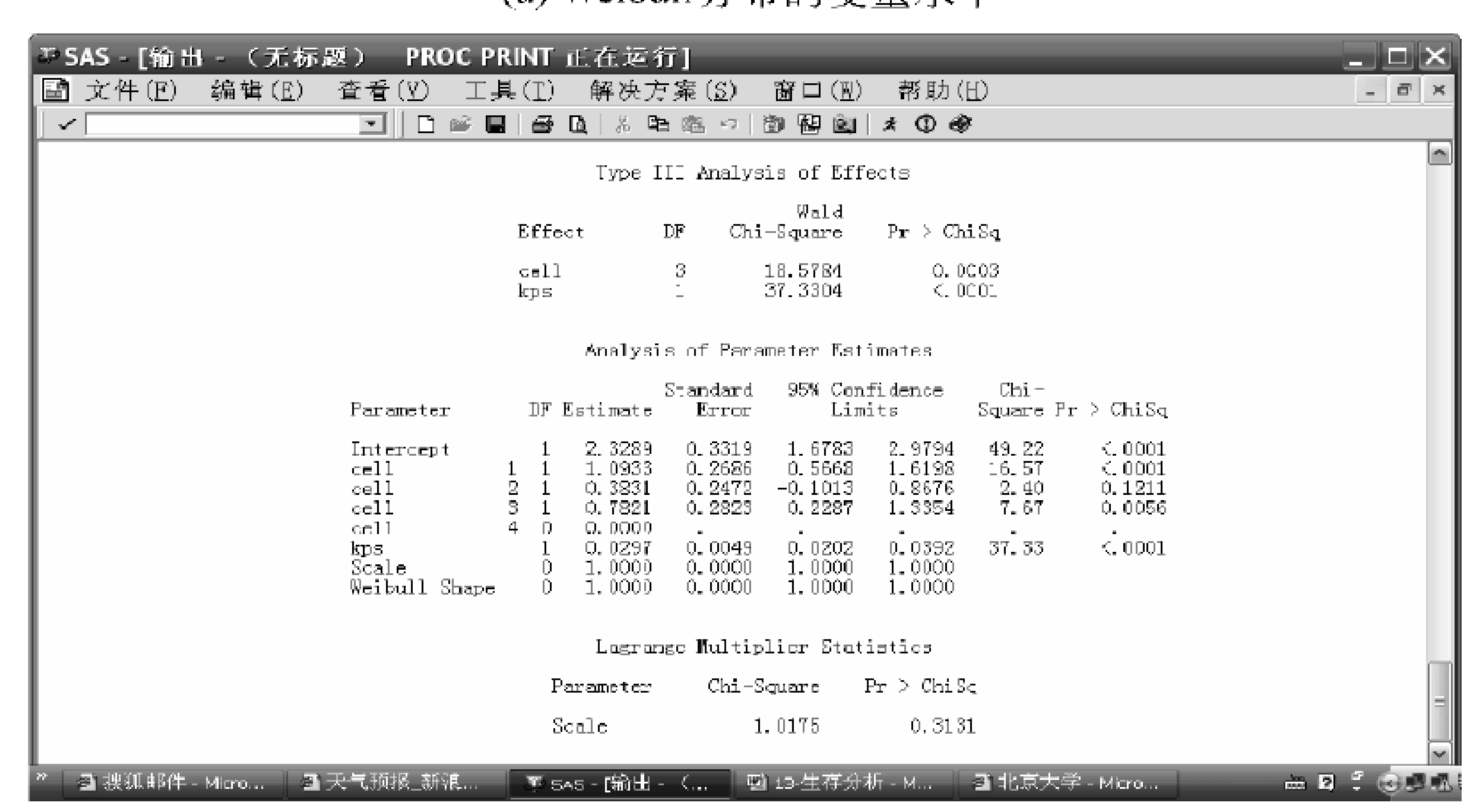

                        Class Level Information

                     Name        Levels    Values

                     therapy          2    1 2
                     cell             4    1 2 3 4


        Algorithm converged.
```

(a) Weibull分布的变量水平



```
SAS - [输出 - (无标题)   PROC PRINT 正在运行]
文件(F) 编辑(E) 查看(V) 工具(T) 解决方案(S) 窗口(W) 帮助(H)

                     Type III Analysis of Effects

                                      Wald
                  Effect      DF    Chi-Square    Pr > ChiSq

                  cell         3       18.5784        0.0003
                  kps          1       37.3304        <.0001


                     Analysis of Parameter Estimates

                                    Standard   95% Confidence     Chi-
      Parameter           DF Estimate   Error      Limits        Square Pr > ChiSq

      Intercept            1   2.3289   0.3319   1.6783   2.9794   49.22   <.0001
      cell             1   1   1.0933   0.2686   0.5668   1.6198   16.57   <.0001
      cell             2   1   0.3831   0.2472  -0.1013   0.8676    2.40   0.1211
      cell             3   1   0.7821   0.2823   0.2287   1.3354    7.67   0.0056
      cell             4   0   0.0000    .        .        .         .       .
      kps                  1   0.0297   0.0049   0.0202   0.0392   37.33   <.0001
      Scale                0   1.0000   0.0000   1.0000   1.0000
      Weibull Shape        0   1.0000   0.0000   1.0000   1.0000

                     Lagrange Multiplier Statistics

                   Parameter      Chi-Square     Pr > ChiSq

                   Scale              1.0175         0.3131
```

(b) Weibull分布的最终结果

图 13.3 拟合基准分布为 Weibull 分布的新结果 1

无标题 - 记事本

文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)

Obs	T	kps	diagtime	age	prior	cell	therapy	CENSOR	_PROB_	PRED
1	72	60	7	69	n	1	1	0	0.5	126.254
2	411	70	5	64	y	1	1	0	0.5	169.931
3	228	60	3	38	n	1	1	0	0.5	126.254
4	126	60	9	63	y	1	1	0	0.5	126.254
5	118	70	11	65	y	1	1	0	0.5	169.931
6	10	20	5	49	n	1	1	0	0.5	38.470
7	82	40	10	69	y	1	1	0	0.5	69.692
8	110	80	29	68	n	1	1	0	0.5	228.720
.....										
25	117	80	3	46	n	2	1	0	0.5	112.427
26	16	30	4	53	y	2	1	0	0.5	25.452
27	151	50	12	69	n	2	1	0	0.5	46.108
28	22	60	4	68	n	2	1	0	0.5	62.060
29	56	80	12	43	y	2	1	0	0.5	112.427
30	21	40	2	55	y	2	1	0	0.5	34.257
31	18	20	15	42	n	2	1	0	0.5	18.910
32	139	80	2	64	n	2	1	0	0.5	112.427
......										
59	278	60	12	63	n	3	1	0	0.5	92.482
60	12	40	12	68	y	3	1	0	0.5	51.050
61	260	80	5	45	n	3	1	0	0.5	167.540
62	200	80	12	41	y	3	1	0	0.5	167.540
63	156	70	2	66	n	3	1	0	0.5	124.477
64	182	90	2	62	n	3	1	1	0.5	225.501
65	143	90	8	60	n	3	1	0	0.5	225.501
66	105	80	11	66	n	3	1	0	0.5	167.540
109	24	40	2	60	n	4	2	0	0.5	23.354
110	18	40	5	69	y	4	2	0	0.5	23.354
111	83	99	3	57	n	4	2	1	0.5	134.783
112	31	80	3	39	n	4	2	0	0.5	76.644
113	51	60	5	62	n	4	2	0	0.5	42.307
114	90	60	22	50	y	4	2	0	0.5	42.307
115	52	60	3	43	n	4	2	0	0.5	42.307
116	73	60	3	70	n	4	2	0	0.5	42.307
......										

图 13.4 拟合基准分布为 Weibull 分布的新结果 2

对图 13.3 结果的分析：

- 数据集的名称为 WORK.Valung。
- 因变量 LOG(T)为追踪寿命。
- 追踪变量为 CENSOR,追踪指示值为 1。
- 非追踪个案 128 个,追踪个案 9 个,而且是右追踪。
- 采用指数最大对数似然估计,其值为−197.55935。
- ChiSquare：卡方值,用于检验"参数为 0"的假设。
- Pr>ChiSq：这是主要的结果,表示 ChiSquare 的概率值(与 α 值比较)。

(1) 模型拟合度

从图 13.3(b)上面看,cell(细胞类型)变量和 kps(手术前的综合体质)变量的 Pr > ChiSq 值都小于 0.0001,因此也都小于 α 值 0.05,非常显著。所以模型拟合度很好。

(2) 生命分析

根据表 13.1,再看图 13.3(b)的下方：cell＝1 为"鱼鳞状"癌细胞,cell＝2 为小细胞,cell＝3为小细胞,cell＝4 为"腺状"癌细胞。由于 cell(细胞类型)有这 4 个水平,便按 1、2、3、4 升序排序,并以第 4 项(ADENO,腺状)项作为基准项(分母),其他项(分子)分别与之比较。可见,cell＝2(小细胞)将是 cell＝4("腺状"癌细胞)生存时间的 $e^{0.3831}$ 倍(1.47 倍)。

(3) 挖掘生存时间

生存回归方程为：

$$T=e^{y}=e^{2.3289+cell*系数+0.0297*kps+Scall*Wp}=e^{2.3289+cell*系数+0.0297*kps+1*Wp}$$

式中,cell 为细胞类型,Wp 为基本分布(如图 13.3 中的 EXPONENTIAL 分布)的 50% 分位数的值。

(4) 当 cell(细胞类型)相同的情况

当 cell 相同时,寿命预测值完全取决于 kps(手术前综合体质评分)。

例如,根据图 13.4 所示的结果,当产生如下值时：

CELL	OBS	KPS	P(PRED)
鱼鳞状(编码为 1)	1	60	126.254
鱼鳞状(编码为 1)	2	70	169.931

则有：Log(P2/P1)＝ Ln(169.931/126.254)＝0.2971＝10 * 0.02971

这个 0.02971 正好是图 13.3(b)中 kps 的 Estimate 值。

(5) kps 值相同但 cell 值不同时的比较：

CELL	KPS	OBS	P(PRED)
鱼鳞状(编码为 1)	60	1	126.254
腺状(编码为 4)	60	113	42.307

则有：Log(P1/P113)＝ Ln(126.254/42.307)＝1.09333

这个 1.09333 正好是图 13.3(b)中 cell＝1(鱼鳞状)的 Estimate 值。

例 7：从某医院随机选择 20 例肺癌患者,记录他们手术后寿命时间及住院号、性别

(变量为 sex)、年龄(age)甲胎蛋白(用 v1 表示,正常值<20ng/ml)、血红蛋白(用 v2 表示,正常值为 110—160g/l)。试用 LIFEREG 过程进行生存回归,并找出影响寿命的因素。

根据数据和题目要求,编辑出以下的程序 13.3(a)。

程序 13.3(a):

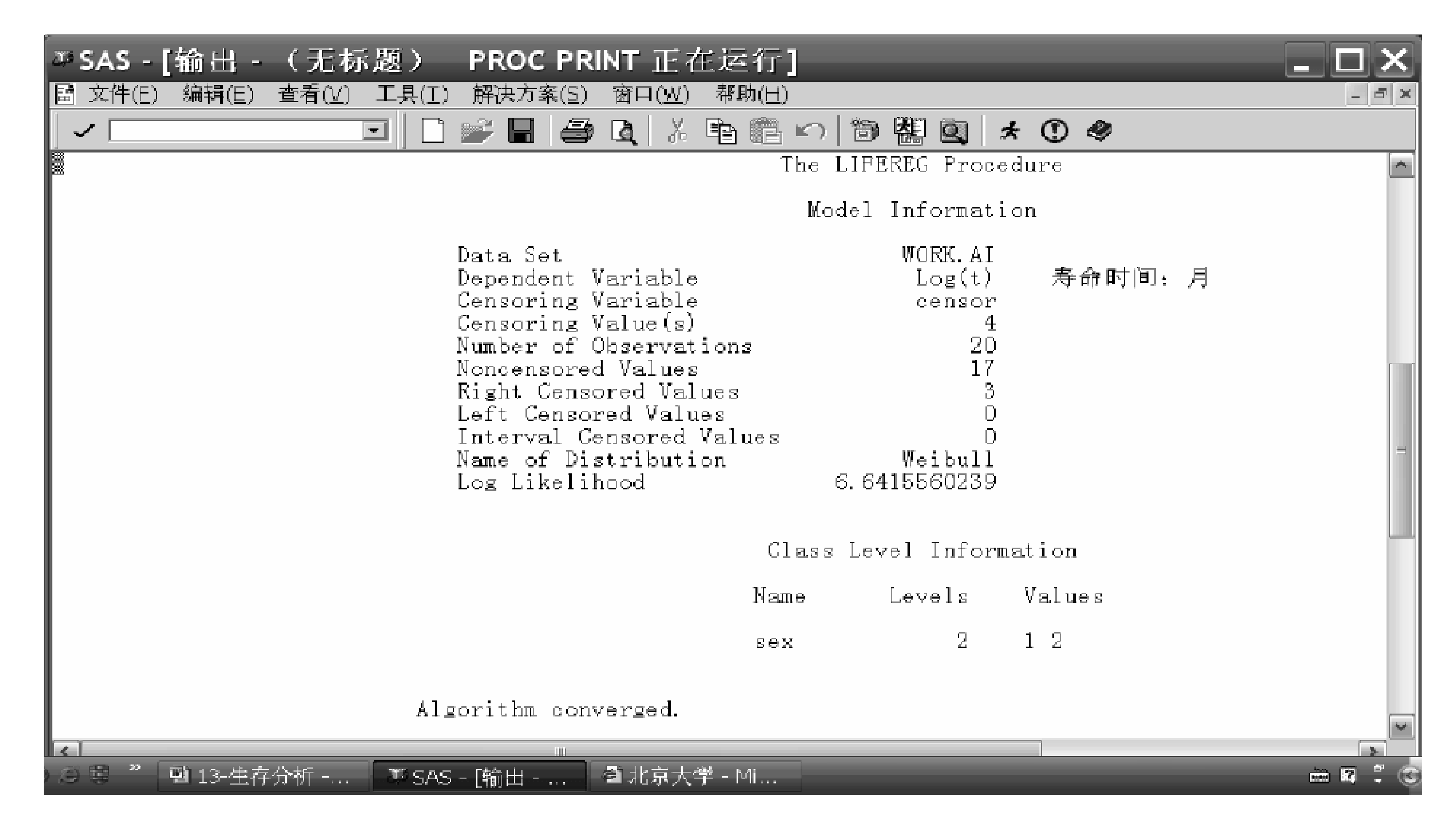
```
DATA ai;
LABEL id = '住院号' sex= '性别' age= '年龄' v1= '甲胎蛋白' v2= '血红蛋白'
     t= '寿命时间:月';
   INPUT id $  sex age opdate mmddyy10. ddate mmddyy10. censor v1 v2 @@ ;
   t= INT((ddate-opdate)/30);                    /* ddate 为 1960.1.1 至去世的天数,opdate 为
                                                    1960.1.1 至手术止的天数 */
CARDS;
3001  1 58  08/20/90  08/21/99  4 20 145 3002 2 60  09/26/91  06/25/98 4 25 160
3103  2 62  01/31/90  03/30/99  2 23 158 3104 2 68  02/28/86  03/27/93 4 26 165
3105  1 65  04/30/85  05/29/96  1 22 161 3206 1 58  05/28/86  05/31/99 1 28 150
3207  1 49  08/28/86  06/28/98  1 25 160 3308 2 64  09/26/84  08/26/97 1 24 158
3309  2 56  08/25/87  08/24/95  2 26 140 3310 1 70  09/28/84  10/30/96 2 18 165
3411  2 73  09/26/85  11/28/95  1 19 162 3512 2 68  08/16/88  08/31/99 1 23 165
3613  2 60  09/26/90  08/28/99  1 25 160 3714 1 58  08/20/90  09/28/99 1 20 163
3715  2 59  09/30/89  09/28/98  2 18 170 3816 1 60  05/31/87  09/30/98 1 24 168
3917  1 50  08/30/90  08/28/99  1 22 162 4018 1 48  07/25/86  08/26/95 2 28 161
4019  1 49  08/24/88  06/05/97  1 26 165 4120 2 60  10/20/89  09/18/94 1 21 160
;
PROC LIFEREG;
   CLASS sex;                                    /* sex 为分类变量 */
  MODEL t * CENSOR(4)= sex age v1 v2/D= WEIBULL;   /* (4)为追踪变量,等号右边为协变量。t 为因
                                                    变量,寿命时间 */
/* 程序 13.3(a)说明:censor 变量是失效(追踪)的指示变量,当 censor= 4 时为失效数据需要追踪和
   检测 */
```

运行程序 13.3(a)产生图 13.5 所示的结果。

从图 13.5(b)看出,sex 和 age 变量的 Pr > ChiSq 值都小于 α 值 0.05,很显著。但 v1 和 v2 变量的 Pr > ChiSq 值都大于 α 值 0.05,不显著,故将 v1 和 v2 变量去掉然后重新建立回归模型如下(见程序 13.3(b))。

程序 13.3(b):

```
DATA ai2;
LABEL id = '住院号' sex= '性别' age= '年龄' v1= '甲胎蛋白' v2= '血红蛋白'
     t= '寿命时间:月';
  INPUT id $  sex age opdate mmddyy8. ddate mmddyy8. censor v1 v2 ;
   t= INT((ddate-opdate)/30);           /* ddate 为 1960.1.1 至去世的天数,opdate 为 1960.1.1 至手术
                                           止的天数 */
```



SAS - [输出 - (无标题) PROC PRINT 正在运行]

The LIFEREG Procedure

Model Information

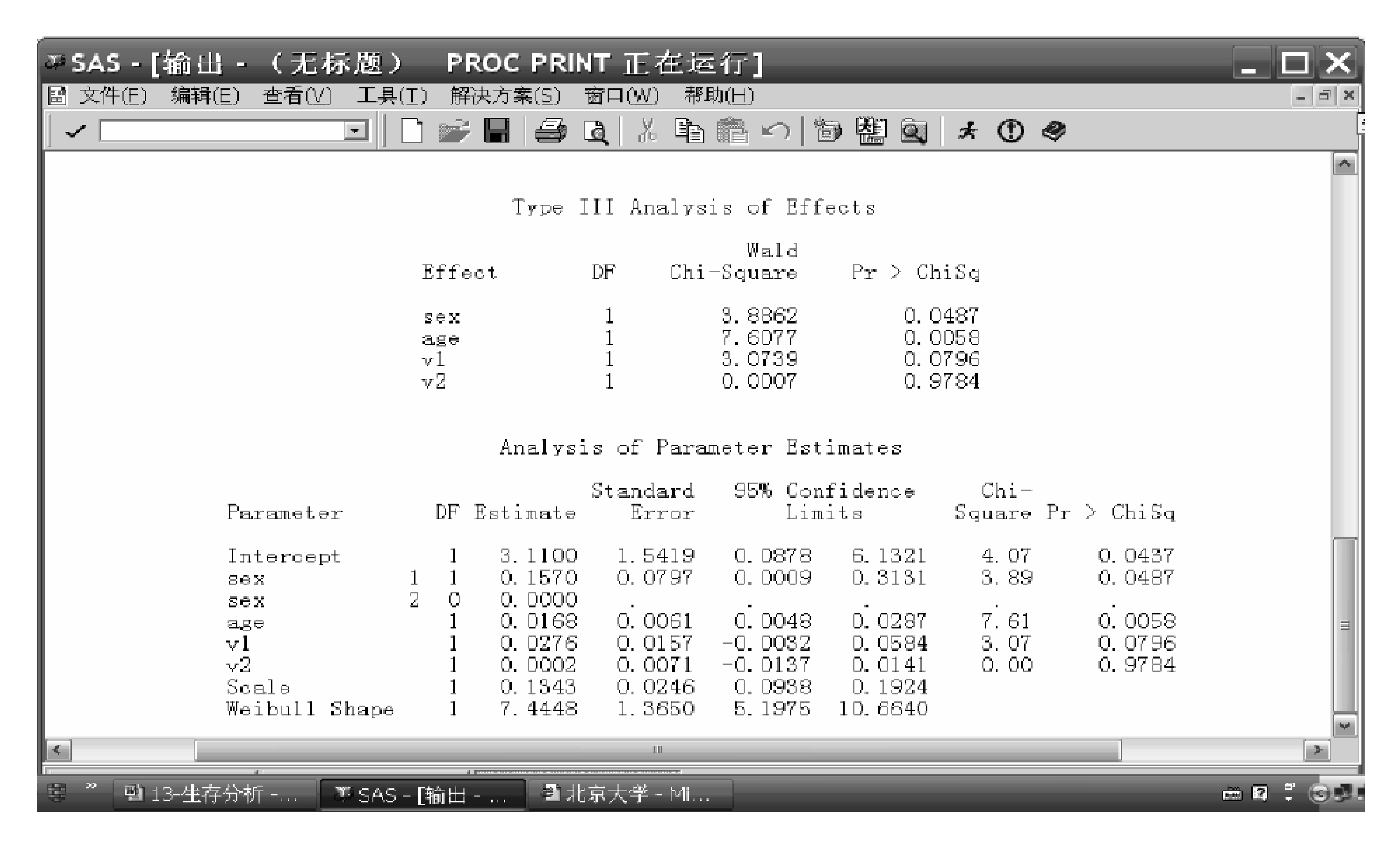
Data Set	WORK.AI	寿命时间：月
Dependent Variable	Log(t)	
Censoring Variable	censor	
Censoring Value(s)	4	
Number of Observations	20	
Noncensored Values	17	
Right Censored Values	3	
Left Censored Values	0	
Interval Censored Values	0	
Name of Distribution	Weibull	
Log Likelihood	6.6415560239	

Class Level Information

Name	Levels	Values
sex	2	1 2

Algorithm converged.

(a) 基本分布为 Weibull 分布的变量水平



SAS - [输出 - (无标题) PROC PRINT 正在运行]

Type III Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
sex	1	3.8862	0.0487
age	1	7.6077	0.0058
v1	1	3.0739	0.0796
v2	1	0.0007	0.9784

Analysis of Parameter Estimates

Parameter		DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept		1	3.1100	1.5419	0.0878	6.1321	4.07	0.0437
sex	1	1	0.1570	0.0797	0.0009	0.3131	3.89	0.0487
sex	2	0	0.0000	.	.	.	.	.
age		1	0.0168	0.0061	0.0048	0.0287	7.61	0.0058
v1		1	0.0276	0.0157	-0.0032	0.0584	3.07	0.0796
v2		1	0.0002	0.0071	-0.0137	0.0141	0.00	0.9784
Scale		1	0.1343	0.0246	0.0938	0.1924		
Weibull Shape		1	7.4448	1.3650	5.1975	10.6640		

(b) Weibull 分布的最终结果

图 13.5 肺癌数据拟合基准分布为 Weibull 分布的结果 1

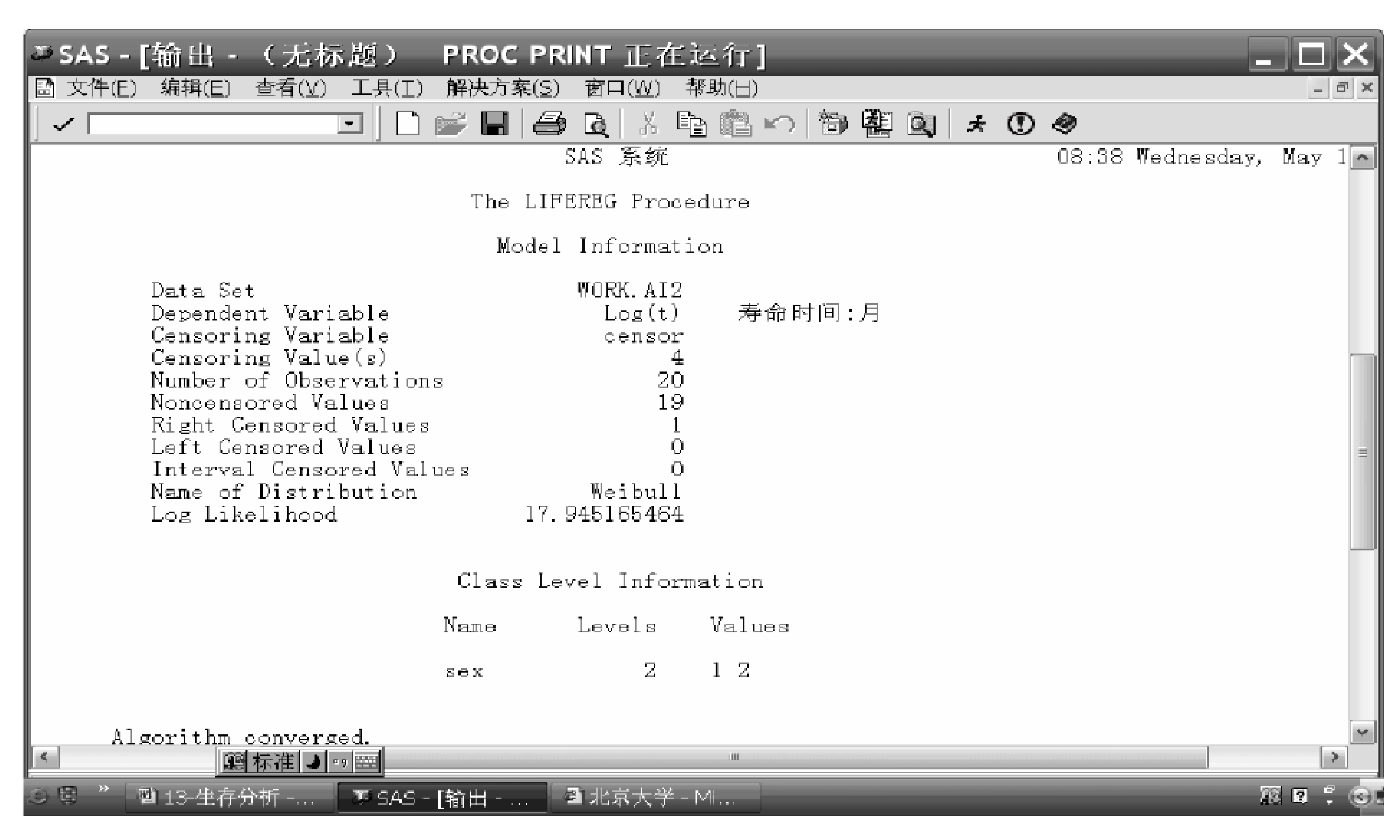
```
CARDS;
3001 1 58 08/20/90  08/21/99 4 20 145
3002 2 60 09/26/91  06/25/98 4 25 160
3103 2 62 01/31/90  03/30/99 2 23 158
3104 2 68 02/28/86  03/27/93 4 26 165
3105 1 65 04/30/85  05/29/96 1 22 161
3206 1 58 05/28/86  05/31/99 1 28 150
3207 1 49 08/28/86  06/28/98 1 25 160
```

```
3308 2 64 09/26/84   08/26/97 1 24 158
3309 2 56 08/25/87   08/24/95 2 26 140
3310 1 70 09/28/84   10/30/96 2 18 165
3411 2 73 09/26/85   11/28/95 1 19 162
3512 2 68 08/16/88   08/31/99 1 23 165
3613 2 60 09/26/90   08/28/99 1 25 160
3714 1 58 08/20/90   09/28/99 1 20 163
3715 2 59 09/30/89   09/28/98 2 18 170
3816 1 60 05/31/87   09/30/98 1 24 168
3917 1 50 08/30/90   08/28/99 1 22 162
4018 1 48 07/25/86   08/26/95 2 28 161
4019 1 49 08/24/88   06/05/97 1 26 165
4120 2 60 10/20/89   09/18/94 1 21 160
;
PROC LIFEREG; CLASS sex;                  /* sex 为分类变量 */
  MODEL t*CENSOR(4)=sex age/D=WEIBULL;
OUTPUT OUT=OUT2 P=PRED;
    PROC PRINT;
```

运行程序 13.3(b)产生图 13.6 至图 13.7 所示的结果。

图 13.6 的结果分析：

图 13.6(c)用于观察原始数据是否顺利地被 SAS 系统统计处理，以及预测的结果如何。这是参考性的图形。



(a) 基本分布为 Weibull 分布的模型概况

图 13.6 基本分布为 Weibull 分布的全部结果

SAS - [输出 - （无标题） PROC PRINT 正在运行]
文件(F) 编辑(E) 查看(V) 工具(T) 解决方案(S) 窗口(W) 帮助(H)

Algorithm converged.

Type III Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
sex	1	1.4148	0.2343
age	1	4.9872	0.0255

Analysis of Parameter Estimates

Parameter		DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept		1	5.2380	0.1690	4.9068	5.5692	960.98	<.0001
sex	1	1	0.0506	0.0426	-0.0328	0.1340	1.41	0.2343
sex	2	0	0.0000	.	.	.	.	.
age		1	0.0058	0.0026	0.0007	0.0108	4.99	0.0255
Scale		1	0.0785	0.0141	0.0552	0.1116		
Weibull Shape		1	12.7388	2.2848	8.9631	18.1050		

13-生存分析 - ... SAS - [输出 - ... 北京大学 - Mi...

(b) 基本分布为 Weibull 分布的模型检验

SAS - [输出 - （无标题） PROC PRINT 正在运行]
文件(F) 编辑(E) 查看(V) 工具(T) 解决方案(S) 窗口(W) 帮助(H)

Obs	id	sex	age	v1	v2	t	opdate	ddate	censor	_PROB_	PRED
1	3001	1	58	4	20	231	11189	18130	9	0.5	269.061
2	3002	2	60	4	25	216	11591	18073	8	0.5	258.751
3	3103	2	62	2	23	233	10988	17986	9	0.5	261.758
4	3104	2	68	4	26	280	9555	17983	3	0.5	270.990
5	3105	1	65	1	22	293	9251	18046	6	0.5	280.164
6	3206	1	58	1	28	280	9644	18048	9	0.5	269.061
7	3207	1	49	1	25	278	9736	18076	8	0.5	255.429
8	3308	2	64	1	24	303	9035	18135	7	0.5	264.800
9	3309	2	56	2	26	267	10098	18133	5	0.5	252.840
10	3310	1	70	2	18	305	9037	18200	6	0.5	288.375
11	3411	2	73	1	19	294	9400	18229	5	0.5	278.931
12	3512	2	68	1	23	256	10455	18140	9	0.5	270.990
13	3613	2	60	1	25	230	11226	18137	9	0.5	258.751
14	3714	1	58	1	20	232	11189	18168	9	0.5	269.061
15	3715	2	59	2	18	243	10865	18168	8	0.5	257.260
16	3816	1	60	1	24	271	10012	18170	8	0.5	272.188
17	3917	1	50	1	22	231	11199	18137	9	0.5	256.909
18	4018	1	48	2	28	281	9702	18135	5	0.5	253.958
19	4019	1	49	1	26	253	10463	18053	7	0.5	255.429
20	4120	2	60	1	21	242	10885	18158	4	0.5	258.751

13-生存分析 - ... SAS - [输出 - ... 北京大学 - Mi...

(c) 基本分布为 Weibull 分布的个案及统计量

图 13.6（续）

从图 13.6(a)看出，本批数据共有 20 个个案(注意，如果此时显示的个案数目不符，则说明数据定义出错，必须先对 INPUT 语句加以调试)，censor ＝4 的个案为追踪的观测值(Censoring Value)。

从图 13.6(b)看出：修改后的程序及输出结果，与图 13.5(b)不同，而且 sex 变量由原来的显著变成了不显著（“Pr＞ChiSq”值 0.2343＞α 值 0.05）。只有 age 变量与生存时间有关。

生存时间 $T = e^{y} = e^{5.238+0.0058*age+SCALL*Wp} = e^{5.238+0.0058*age+0.0785*Wp}$

可简化为 $T = e^{5.238+0.0058*age}$

13.3 用 LIFETEST 过程进行生存检验

LIFETEST 过程可对有追踪(未失效)的数据进行生存分布的非参数估计,以及计算因变量与自变量相关性的秩次检验结果。非参数估计是分层进行的,秩次检验是合并的。另外还输出"检验各层之间一致性"的统计量。因此,LIFETEST 过程是生存检验过程。

13.3.1 生存分布函数 SDF 及其他函数

对有追踪(未失效)的数据进行非参数估计的第一步,是估计寿命分布,因此首先要知道生存分布函数(即 SDF)在 T_i 处的值。

1. 用 LIFETEST 过程进行非参数估计的步骤

(1) 若用 SDF 表示来自总体的某随机样本,则寿命分布在 T_i 处的值为:

$$S(t) = \text{Prob}(T > t) \tag{13.7}$$

其中,"$T>t$"是条件成立的概率。

(2) 再用极限乘法计算寿命分布函数

用极限乘法(product limit estimate)计算寿命分布函数的公式如下:

$$S(T_i) = \prod_{j}^{i} (1 - D_i/N_j) \tag{13.8}$$

式(13.8)中,i 为 1、2、3、…、n。T_i 是不同的追踪时间,D_i 是在 T_i 时间点去世的个体数目,N_j 是在 T_i 时间点追踪的个体数目。

2. 其他函数

(1) 累积分布函数(CDF)$F(t)$的表示法:

$$F(t) = 1 - S(t) \tag{13.9}$$

其中,$1-S(t)$为死亡概率或失效概率。

(2) 概率密度函数

$f(t)=F(t)$的导数。

(3) 危险率(Hazard)函数的表示法:

$$H(t) = f(t)/F(t) \tag{13.10}$$

13.3.2 LIFETEST 过程的命令语句

1. LIFETEST(生存检验过程)的命令语句

```
PROC LIFETEST[选项];          /* 过程名,主语句。用于调用后面的语句和数据。必须有的
                                语句 * /
TIME t * CENSOR(1).;          /* 必须有的语句,其中 t是寿命时间变量(即因变量),CENSOR(1)
```

```
                              表示指示变量是 CENSOR,CENSOR= 1为追踪值 * /
STRATA v1 v2(范围);           /* v1、v2等表示变量表,定义分层的变量, * /
  TEST v1 v2 v3;              /* TEST语句中的变量是被检验的变量,如 v1、v2、v3等变量 * /
  ID v1;                      /* ID语句对输出结果起作用,即用 v1变量(其他变量也行)标识寿命估
                                 计时的个案 * /
FREQ v;                       /* 指定频数做统计的变量 * /
  BY v;
```

2. 选项

选项是指以下内容:

- DATA=数据集名称
- METHOD=PL|KM|LT|LIFE|ACT

(其中,PL 或 KM 为极限乘法。LT 或 LIFE 或 ACT 为寿命表法。)

- MISSING /* 将数字型变量的缺失值(或字符型变量中的空格)作为有效值 */
- OUTTEST=out1 /* 指定输出数据集名称为 out1。那么,秩次检验的统计量、协方差矩阵、层间的各种卡方检验值便可自动存入 out1 文件中 */
- PLOTS=(散点图名称)

例如:PLOTS=(S,LLS),表示产生 S 和 LLS 两个散点图

PLOTS=(S) 或 PLOTS=(SURVIVAL),表示产生 S 寿命函数图

PLOTS=(LS) 或 PLOTS=(LOGSURVIVAL),代表-LOG(SDF)

PLOTS=(LLS) 或 PLOTS=(LOGLOGS),代表 LOG(-LOG(SDF))

PLOTS=(H) 或 PLOTS=(HAZARD),代表危险函数

PLOTS=(P) 或 PLOTS=(PDF),代表概率密度函数

- INTERVALS=(数值表) /* 为寿命表法指定区间终点,初始值一定为 0。用于计算寿命分布函数 */
- WIDTH = 值 /* 为寿命表法计算寿命分布时指定区间宽度。若指定了 INTERVALS=项,则此项无效 */
- INTERVAL =值 /* 为寿命表法指定区间数,默认为 0。若指定了INTERVALS=(数值表)或 WIDTH=,则此项无效 */
- ALPHA=值 /* 计算寿命估计的置信区间的置信水平,0.0001< 值<0.9999。置信水平为 1- ALPHA,默认为 ALPHA =0.05,置信水平为 95% */

3. TIME 语句

TIME t * CENSOR(1,2); /* t 是寿命时间变量(即因变量),当 CENSOR=1 或 CENSOR=2 时表示追踪值(失访) */

4. STRATA 语句

STRATA age(10,20 TO 60 BY 10); /* age 是分层变量,有[0,10]、[10,20]、[20,30]、[30,40]、[40,50]、[50,60]、[60,以上]一共 7 个水平 */

5. TEST 语句

TEST age score / * 列出数字型自变量，以便检验这些自变量与因变量的关联度 * /

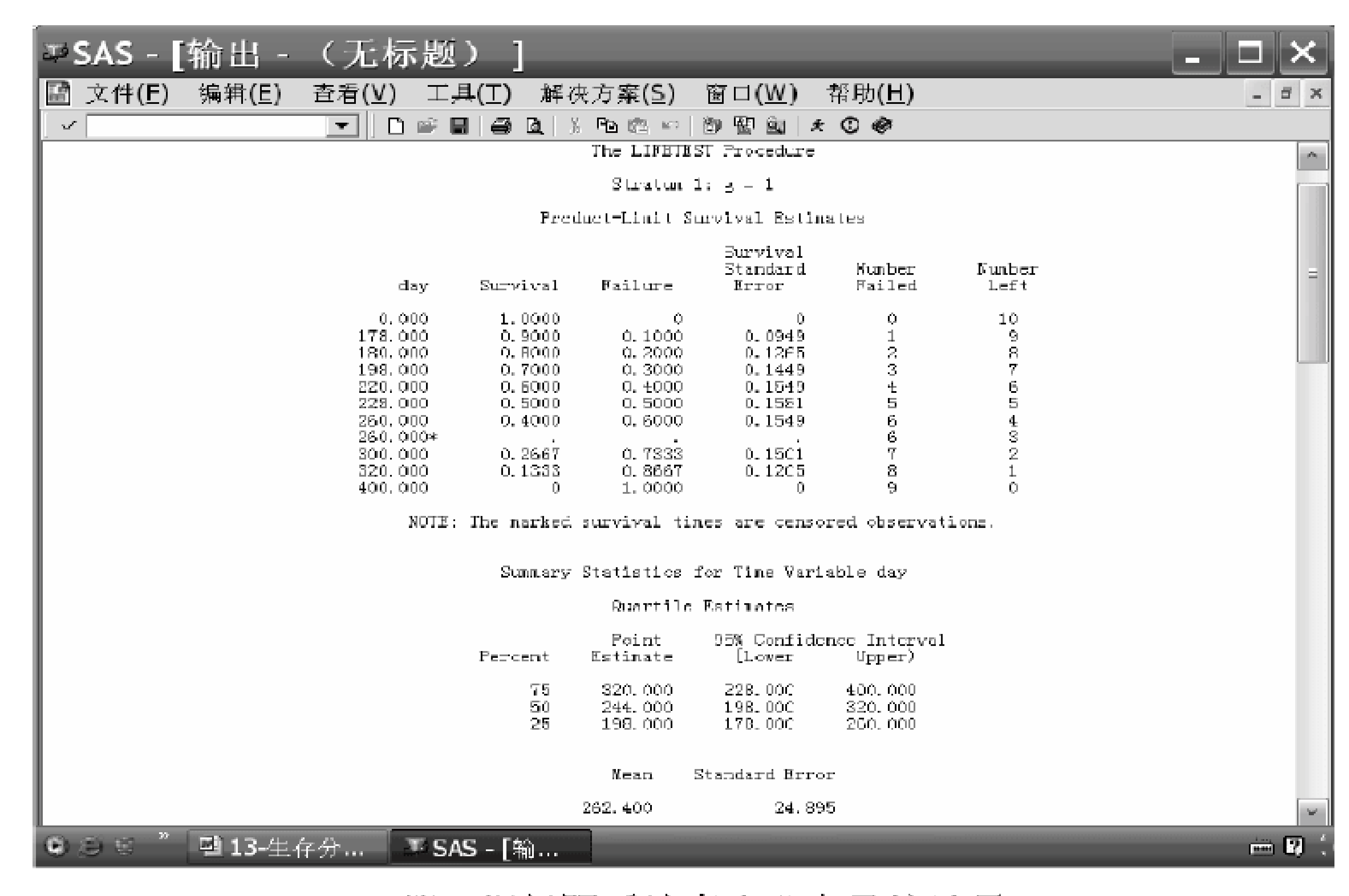
13.3.3 应用举例

例 8：这里引用某医院手术的一个案例说明生存分布函数的计算法。数据和命令见程序 13.4。

程序 13.4：

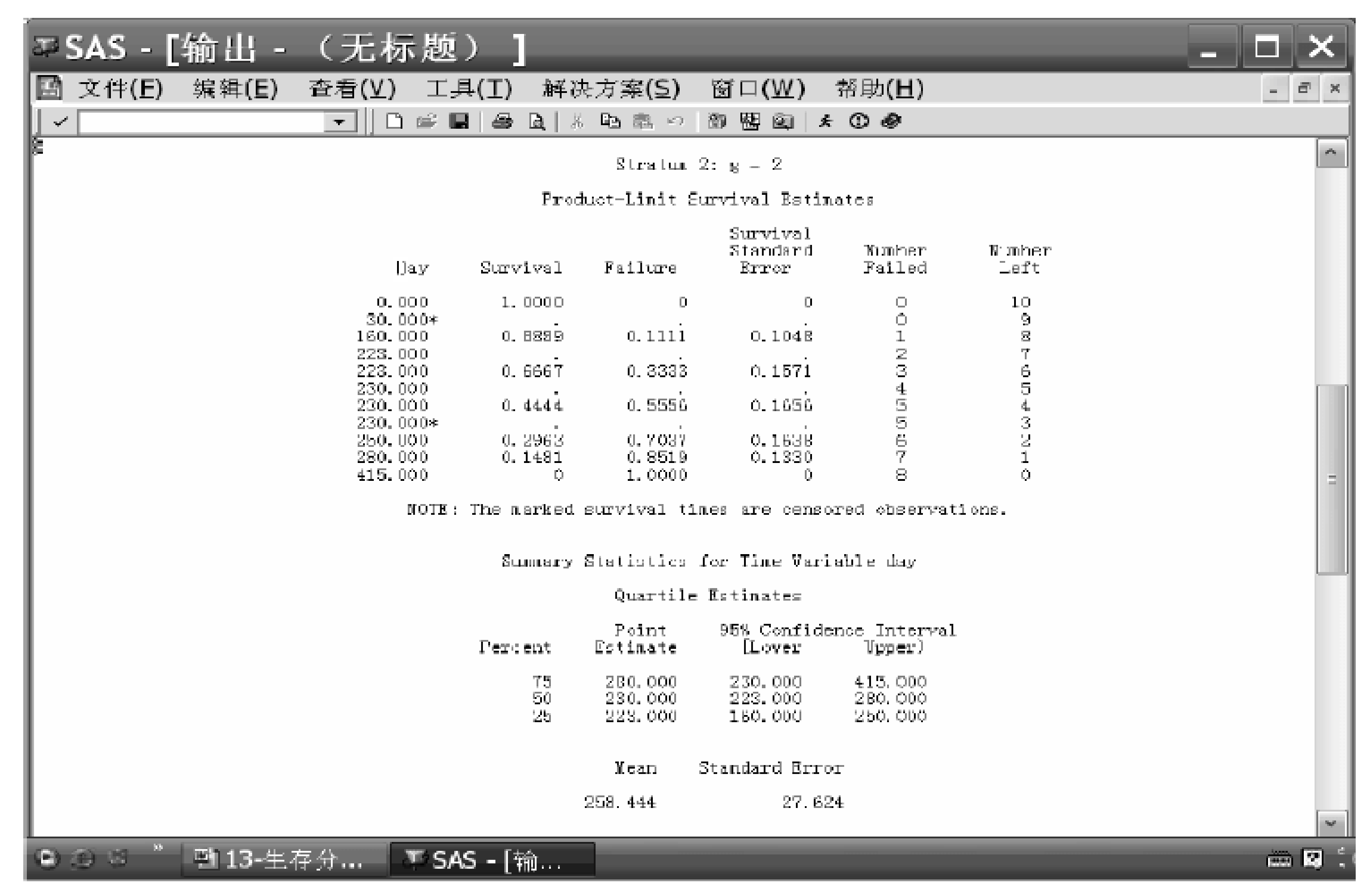
```
TITLE '手术后的生存分布';
DATA AI2;
LABEL day='从手术到死亡的天数'   g='个案分组。前 10 人一组，其余为另一组。';
INPUT day @@;
CENSOR=day<0;                        /* 寿命时间为负时是追踪个案 * /
     day=ABS(day);                   /* 因变量取正值 * /
     g=(_N_ >10)+1;                  /* 前 10 人一组，其余为另一组 * /
CARDS;
228 320 400 300 180 220 178 260 198 -260
-230 -30 415 280 160 250 230 230 223 223
;
PROC LIFETEST PLOTS=(s,LLS);         /* 产生生存分布函数及散点图，* /
TIME day*CENSOR(1);                  /* 当 CENSOR=1 时为追踪个案 * /
STRATA g;                            /* g 为分层变量，以便比较两层之间的生存分布的
                                        一致性 * /
RUN;
```

运行程序 13.4 产生图 13.7 所示的结果。



(a) 第一组极限乘法寿命分布及统计量

图 13.7 两组手术患者的寿命估计



Stratum 2: g = 2

Product-Limit Survival Estimates

Day	Survival	Failure	Survival Standard Error	Number Failed	Number Left
0.000	1.0000	0	0	0	10
30.000*	.	.	.	0	9
160.000	0.8889	0.1111	0.1048	1	8
223.000	.	.	.	2	7
223.000	0.6667	0.3333	0.1571	3	6
230.000	.	.	.	4	5
230.000	0.4444	0.5556	0.1656	5	4
230.000*	.	.	.	5	3
250.000	0.2963	0.7037	0.1538	6	2
280.000	0.1481	0.8519	0.1330	7	1
415.000	0	1.0000	0	8	0

NOTE: The marked survival times are censored observations.

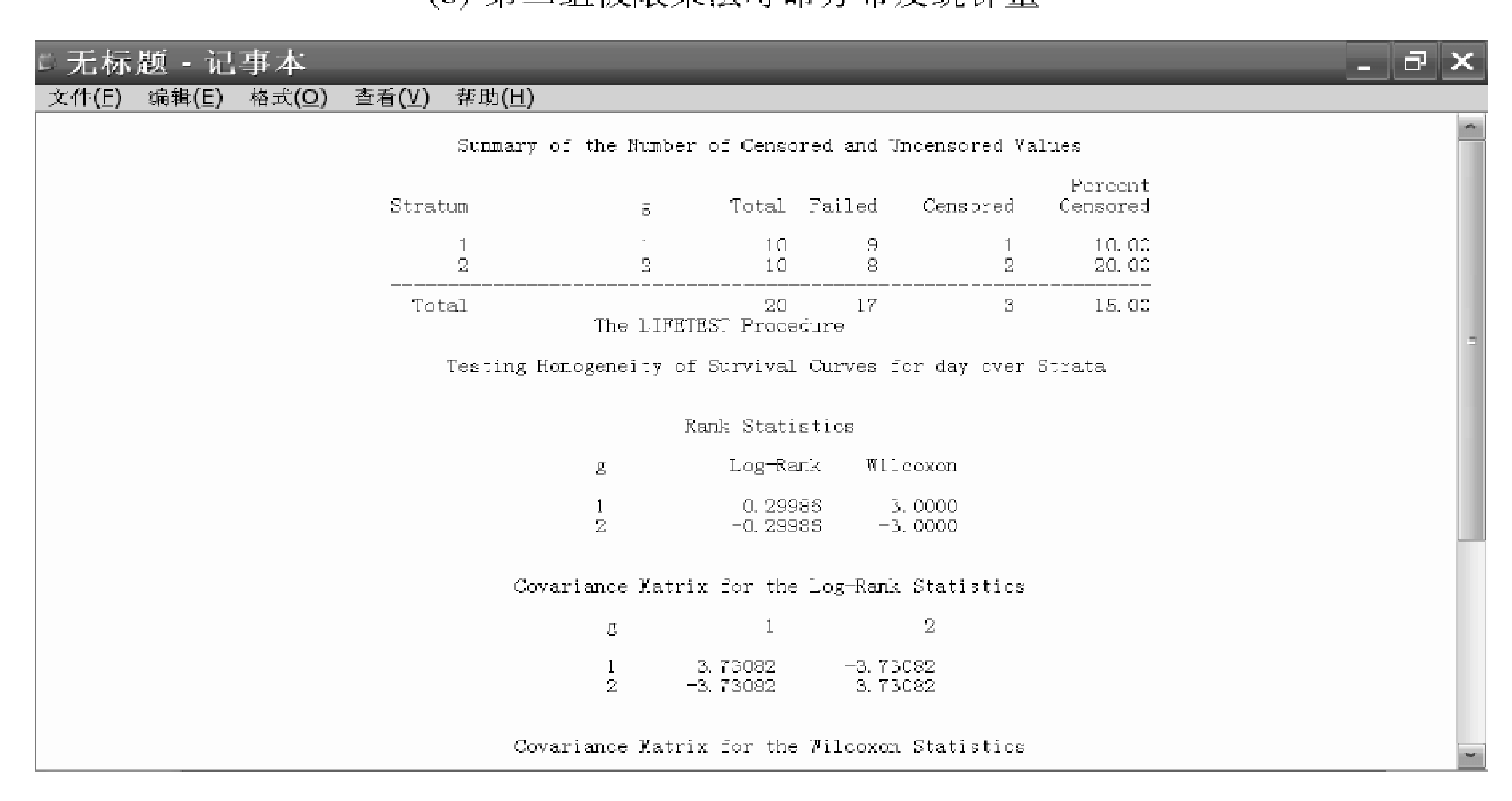
Summary Statistics for Time Variable day

Quartile Estimates

Percent	Point Estimate	95% Confidence Interval [Lower	Upper)
75	280.000	230.000	415.000
50	230.000	223.000	280.000
25	223.000	160.000	250.000

Mean	Standard Error
258.444	27.624

(b) 第二组极限乘法寿命分布及统计量



Summary of the Number of Censored and Uncensored Values

Stratum	g	Total	Failed	Censored	Percent Censored
1	1	10	9	1	10.00
2	2	10	8	2	20.00
Total		20	17	3	15.00

The LIFETEST Procedure

Testing Homogeneity of Survival Curves for day over Strata

Rank Statistics

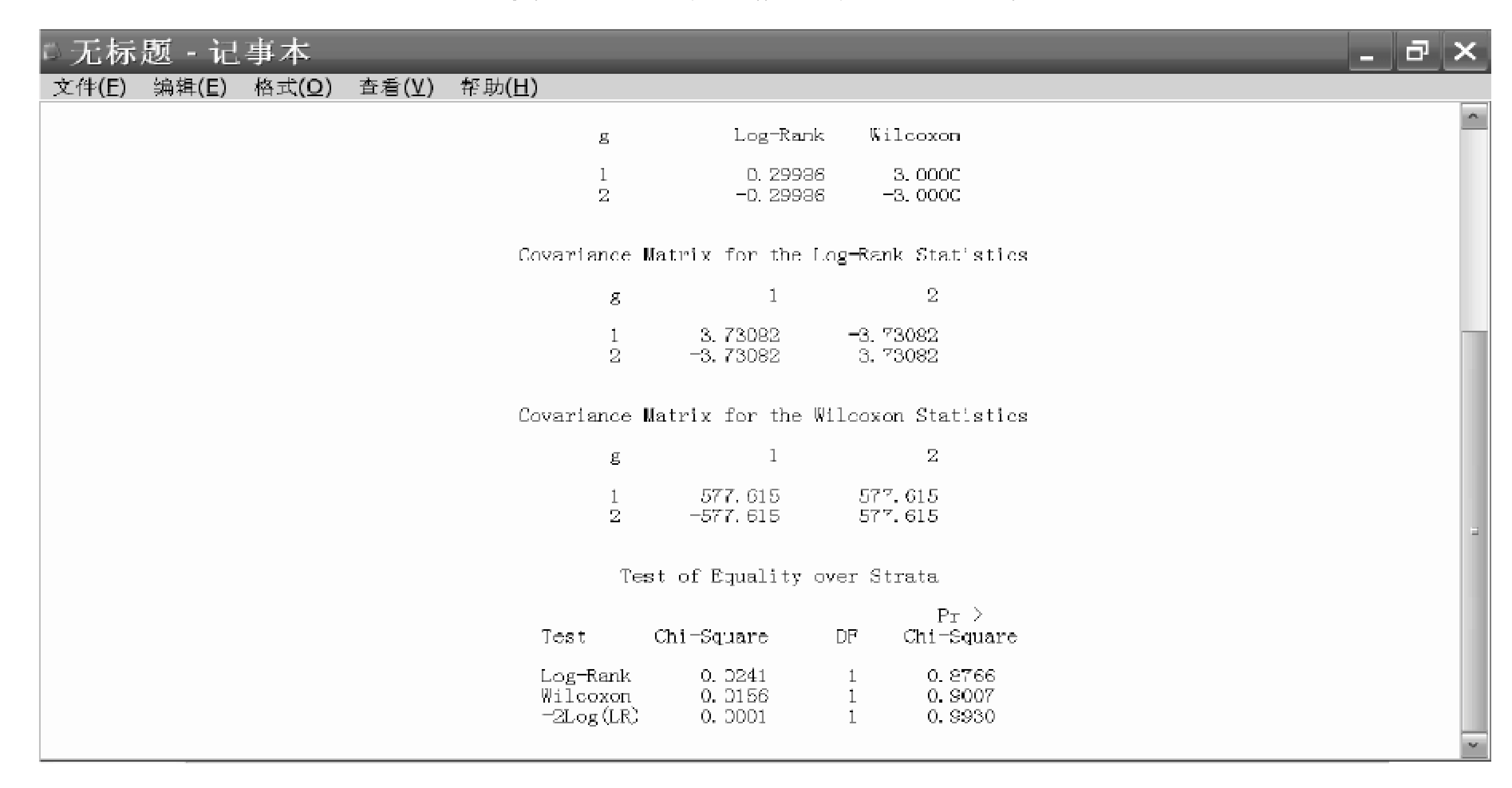
g	Log-Rank	Wilcoxon
1	0.29986	3.0000
2	-0.29986	-3.0000

Covariance Matrix for the Log-Rank Statistics

g	1	2
1	3.73082	-3.73082
2	-3.73082	3.73082

Covariance Matrix for the Wilcoxon Statistics

(c) 两组追踪与非追踪的个案比较



g	Log-Rank	Wilcoxon
1	0.29986	3.0000
2	-0.29986	-3.0000

Covariance Matrix for the Log-Rank Statistics

g	1	2
1	3.73082	-3.73082
2	-3.73082	3.73082

Covariance Matrix for the Wilcoxon Statistics

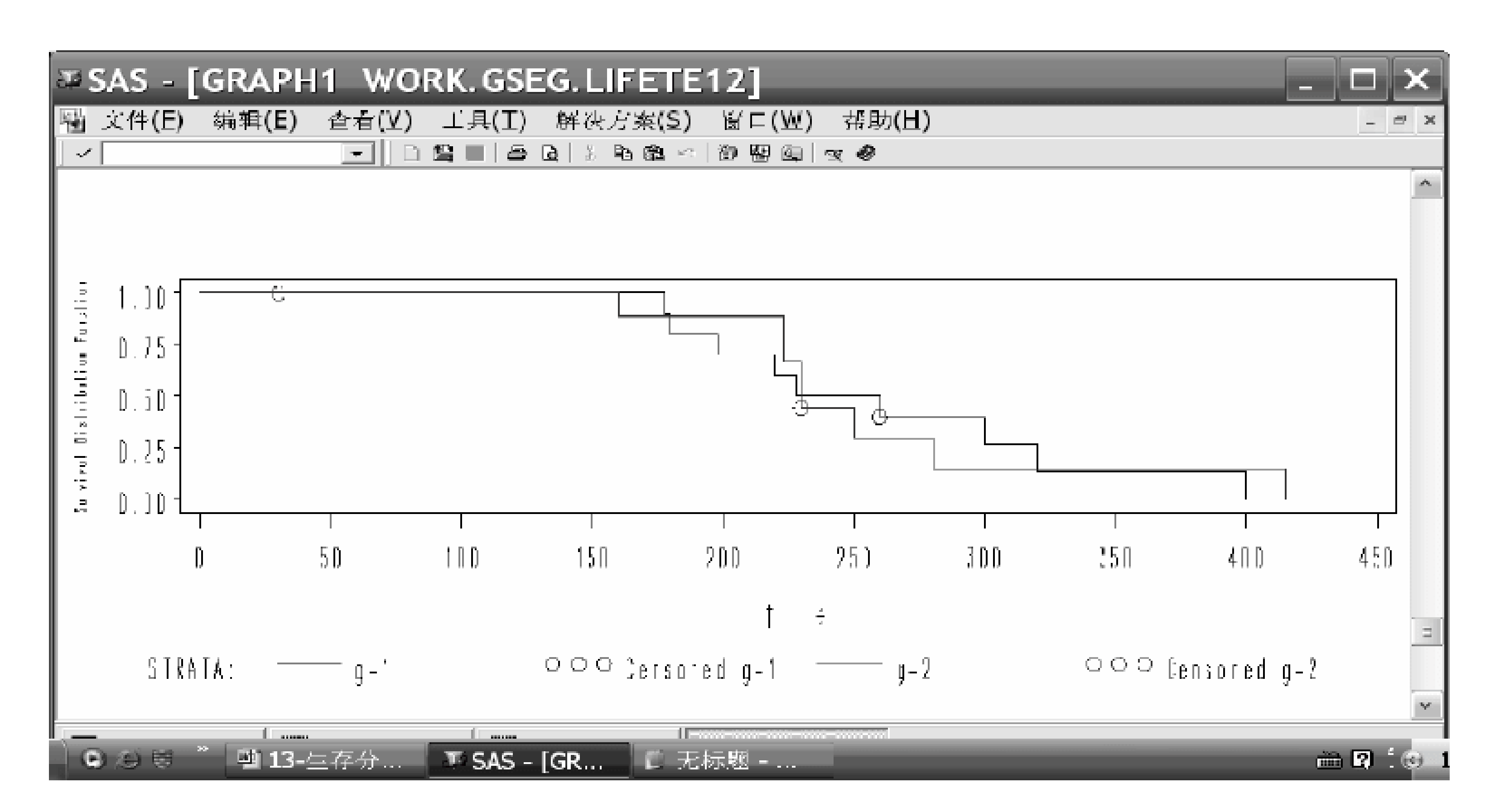
g	1	2
1	577.615	577.615
2	-577.615	577.615

Test of Equality over Strata

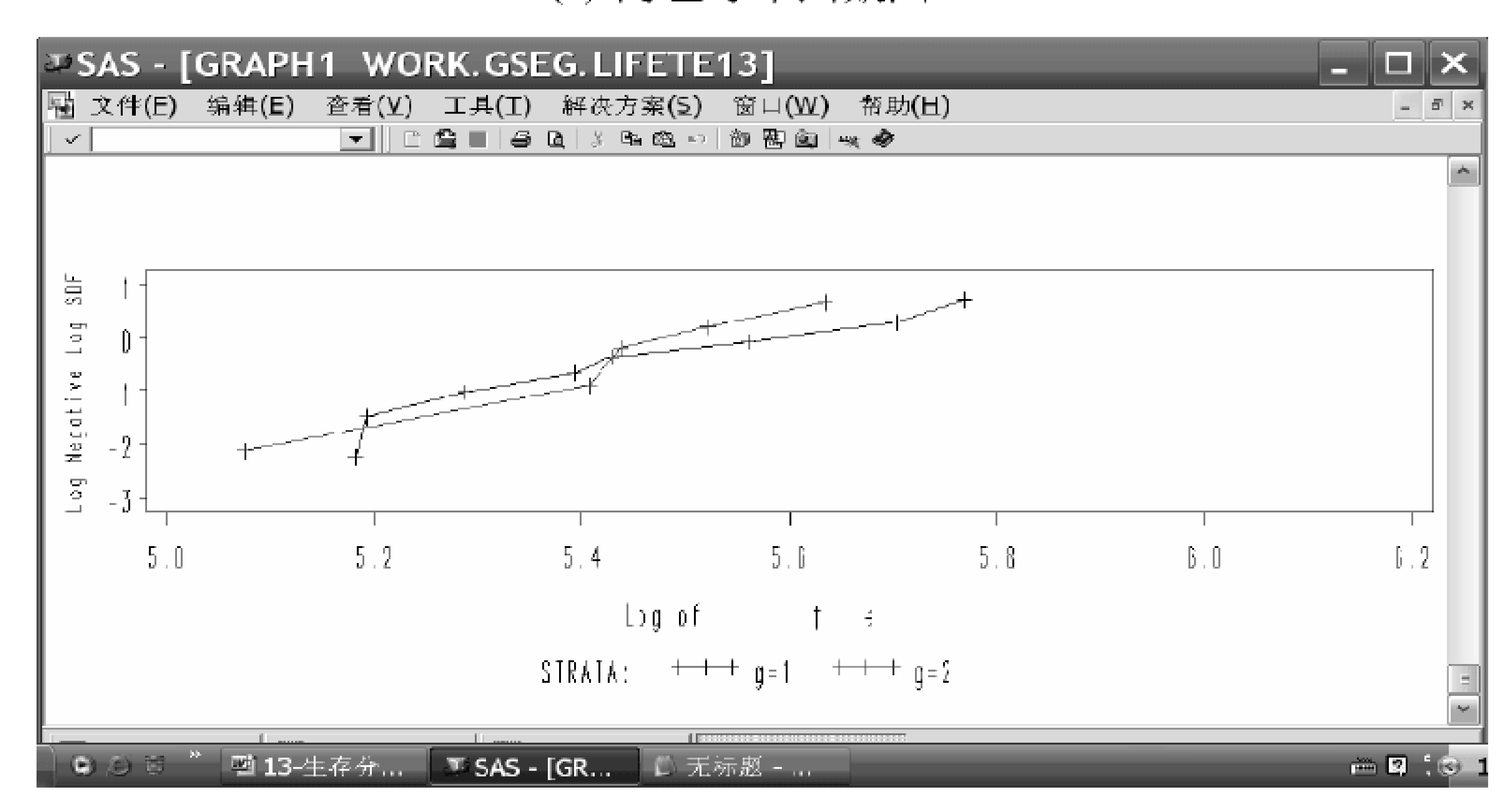
Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	0.0241	1	0.8766
Wilcoxon	0.0156	1	0.9007
-2Log(LR)	0.0001	1	0.9930

(d) 各层之间一致性检验

图 13.7（续）



(e) 两组寿命函数图



(f) 两组对数寿命函数图

图 13.7（续）

图 13.7(b)中的主要统计量名称解释如下：

Day　因变量，患者从手术到死亡的天数

Survival　寿命概率

Failure　失效(或死亡)的概率＝1－Survival，见图 13.7(b)

Survival Standard Error　标准误差

Number Failet　非追踪(已去世)的人数

Number Left　追踪(生存着)的人数

Quartile Estimate　因变量四分位数估计

Mean　因变量的均值

Point Estimate　点估计值

从图 13.7 的(c)图和(d)图看，由于过程命令使用了“STRARA g”语句，因此生存估计是按照变量 g 的两个值分层比较两层之间的一致性。读者还可以从图 13.7 的(b)～(d)几个子图中了解每一层内某一天所对应的生存概率、死亡概率、标准误差、生存人数、死亡人数，以及生存天数的四分位数分析及均值、标准误差等。

而且，从图 13.7 的(d)图和(c)图看，用 Log-Rank 检验时，其“Pr>Chi-Square”值0.8766大于 α 值 0.05，表明两组寿命基本相同。同理，用 Wilcoxon 等其他两种方法检验时，两组寿命也基本相同。

图 13.7(e)和(f)是散点图，对应于“PLOTS=(s，LLS)”语句，画的是生存分布函数曲线。纵坐标为生存概率，横坐标为生存天数。由于选择了“STRARA g”语句，所以显示两条曲线。从图 13.7(e)和(f)看，左侧较长的线表示第一组，另一条线表示第二组。生存时间的初期(约 325 天以前)，两组曲线才有所差别。但当生存时间值变大时，两组曲线就没有差别。

例 9：有一批胃癌患者的年龄、胃癌发现的病期(g=1 为早期，g=2 为中期，g=3 为晚期)及做切除手术后的生存时间(T)见程序 13.5。试用 LIFETEST 过程对程序 13.6 中的数据进行生存估计，要求分析病期与手术后的寿命之间的关系，并做组间寿命比较。

程序 13.5：

```
TITLE '手术后的生存估计';
DATA WA;
LABEL id='个案号'  患者 g='患者的病期'  age='患者的年龄'  t='存活时间:天';
Group=(_N_<7);            /* _N_为内部隐含变量。第 1~6个个案为第 1组,其余为第 2组 */
INPUT id g age t @@;
CARDS;
01 2 68 234 02 1 54 200 03 1 69 380 04 1 50 350
05 3 53 200 06 3 49 190 07 2 70 330 08 3 68 290
09 3 68 134 10 3 54 170 11 1 66 310 12 1 46 270
;
PROC LIFETEST;
STRATA group;
    TIME t;
  TEST age g;             /* g为分层变量,以便比较两层之间的生存分布的一致性 */
RUN;
```

程序 13.5 说明：

Group=(_N_<7)：_N_为系统内部默认的变量。第 1～6 个个案为第 1 组，其余为第 2 组。

TIME t：将原始的存活时间 t 赋予变量 TIME，因变量为 TIME。

TEST age g：指明要检验的协变量为 age 等变量。

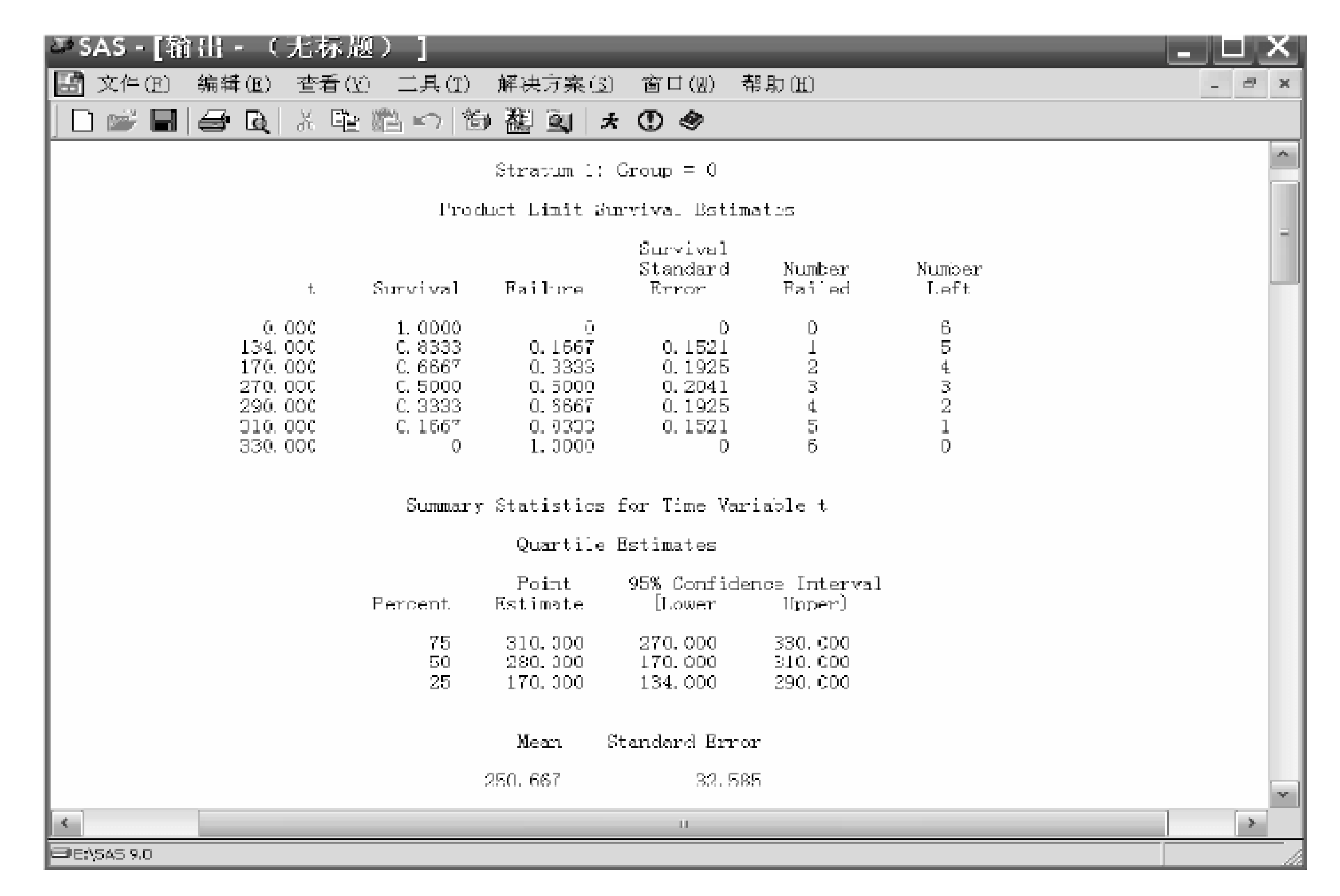
运行程序 13.5 产生图 13.8 所示的结果。

结果分析：

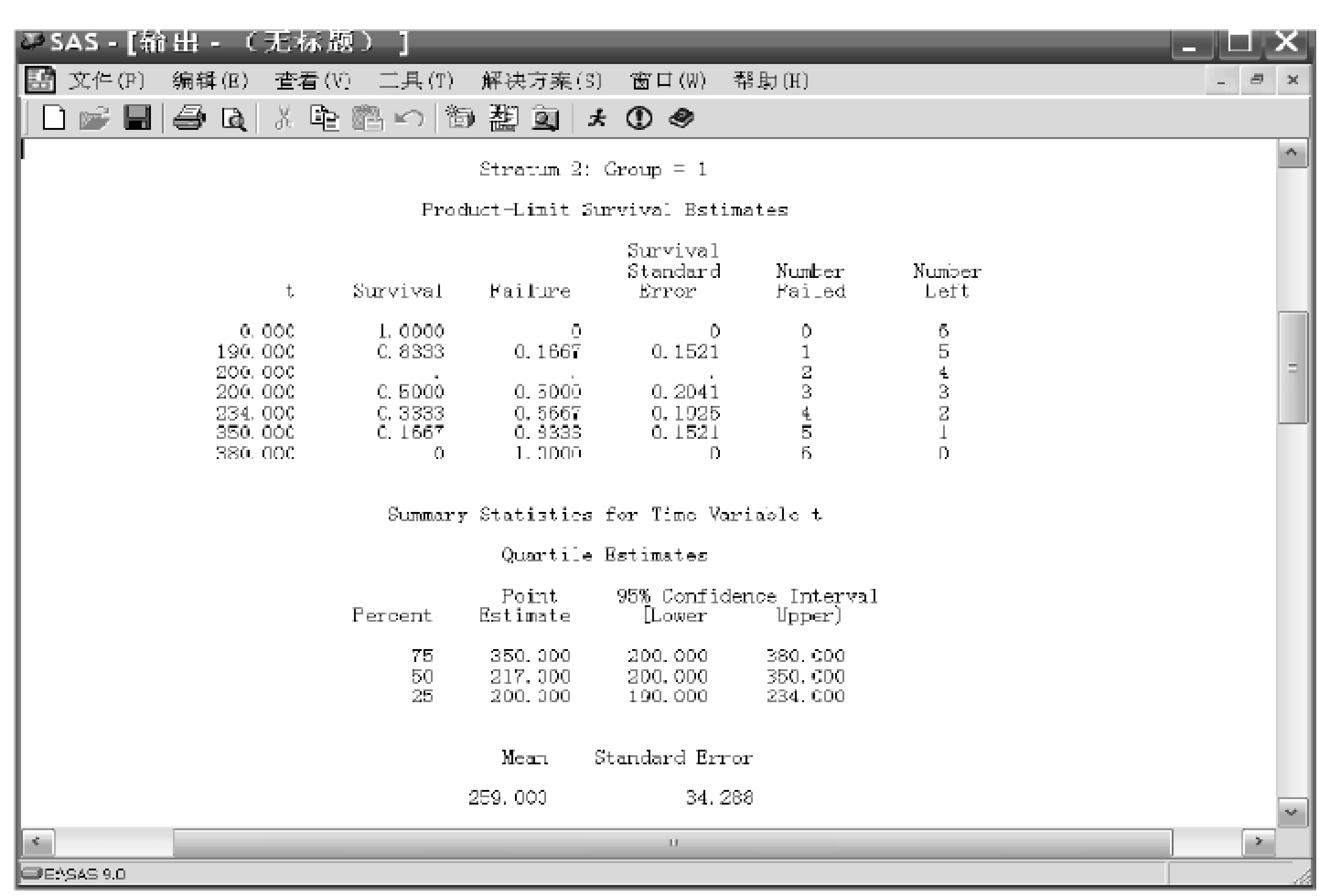
(1) 图 13.8 中的(a)和(b)两个子图是患者分为两组时的寿命率及四分位数寿命率分布，仅供参考。

(2) 图 13.8(c)是各层一致性的三种检验法，由于三种检验的“Pr>Chi-Square”值分别都大于 α 值 0.05，说明两组患者的寿命分布的差别为 0。

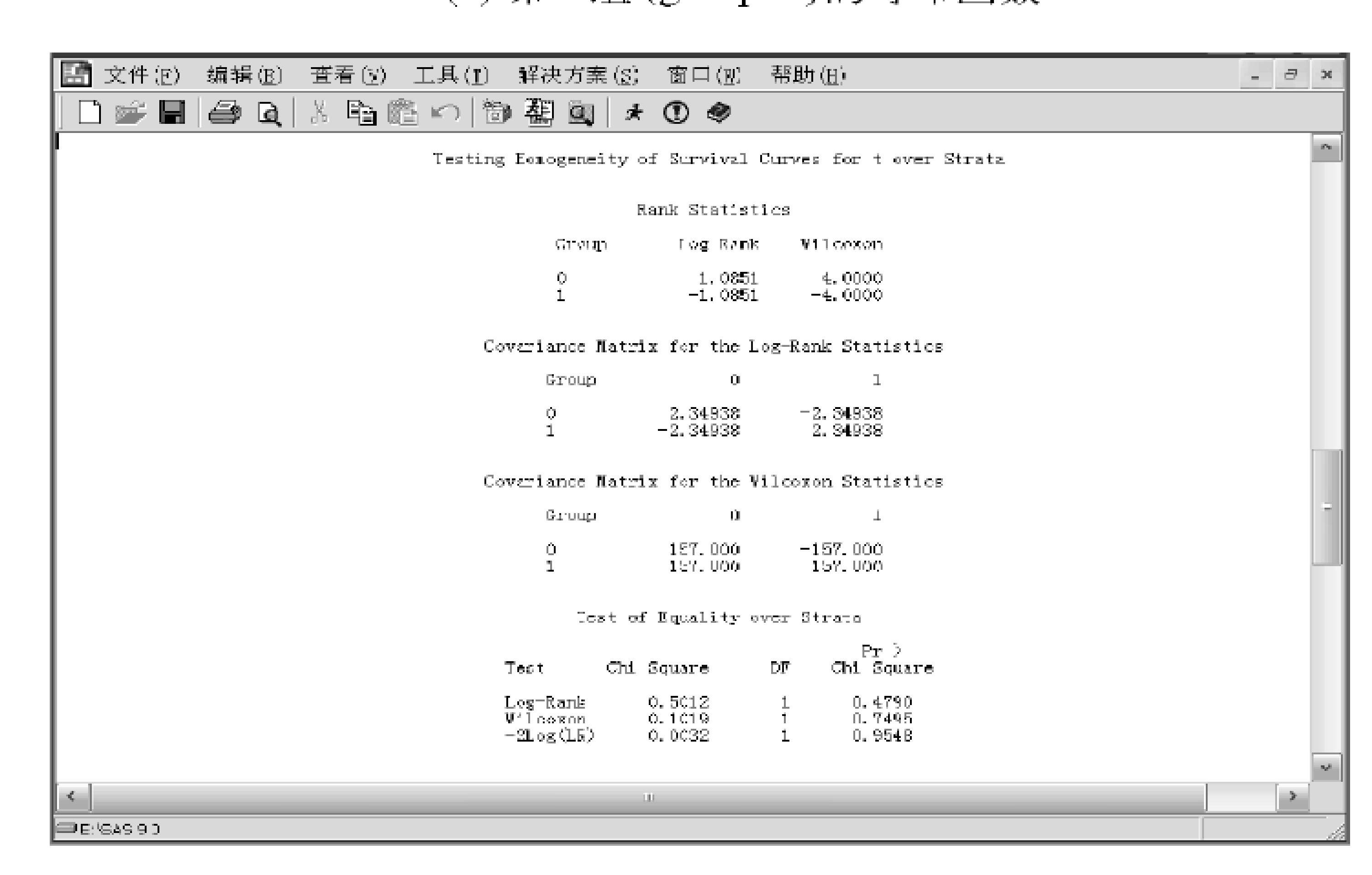
(3) 图 13.8(d)是因变量与协变量相关性(结合度)的 Wilcoxon 检验，主要有以下 3 种检验：



(a) 第 1组 (group=0)的寿命函数

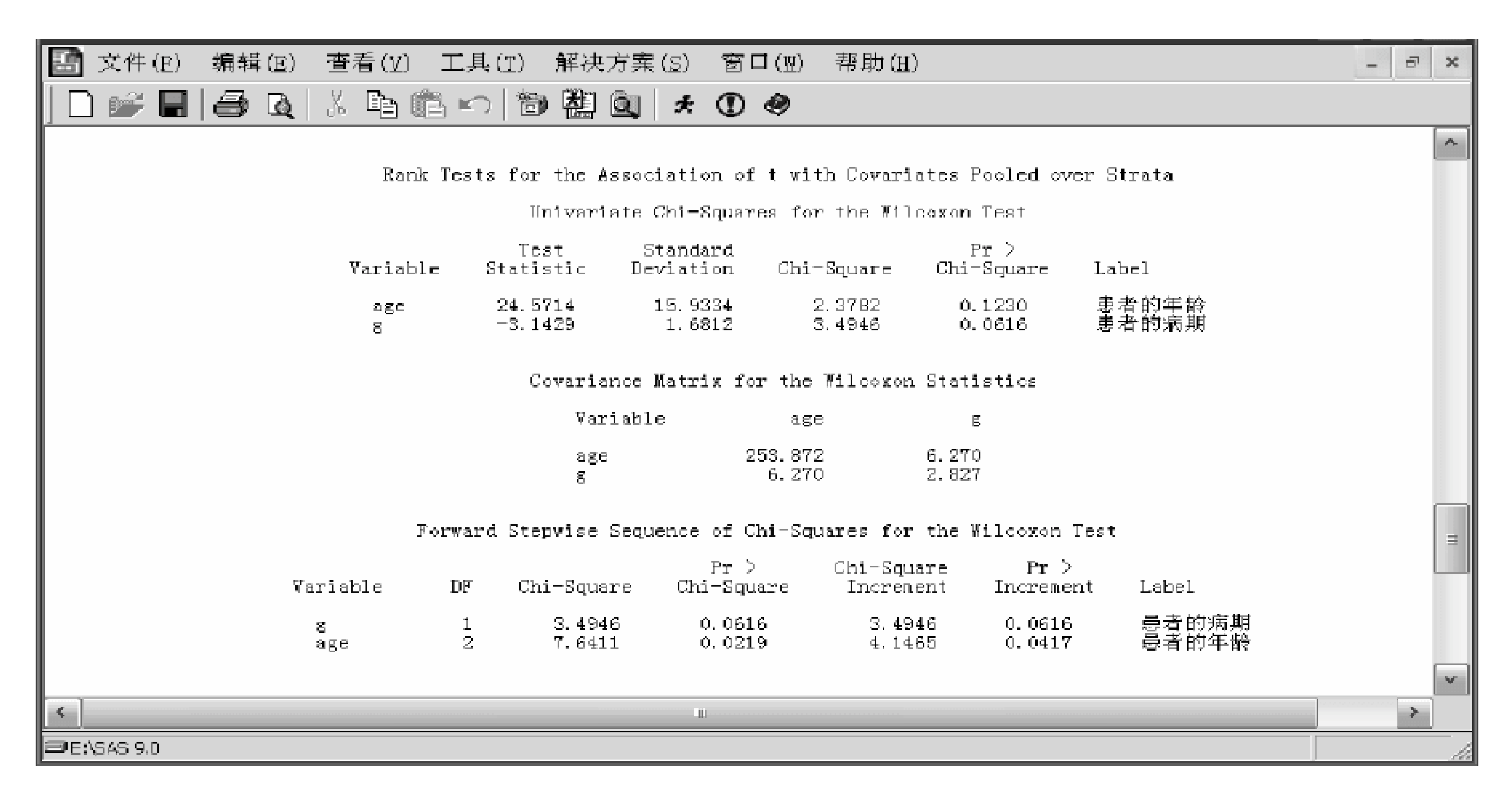


(b) 第 2组 (group=0)的寿命函数



(c) 各层间一致性的 3种检验法

图 13.8　分组的病期与寿命关系分析



Rank Tests for the Association of t with Covariates Pooled over Strata

Univariate Chi-Squares for the Wilcoxon Test

Variable	Test Statistic	Standard Deviation	Chi-Square	Pr > Chi-Square	Label
age	24.5714	15.9334	2.3782	0.1230	患者的年龄
g	-3.1429	1.6812	3.4946	0.0616	患者的病期

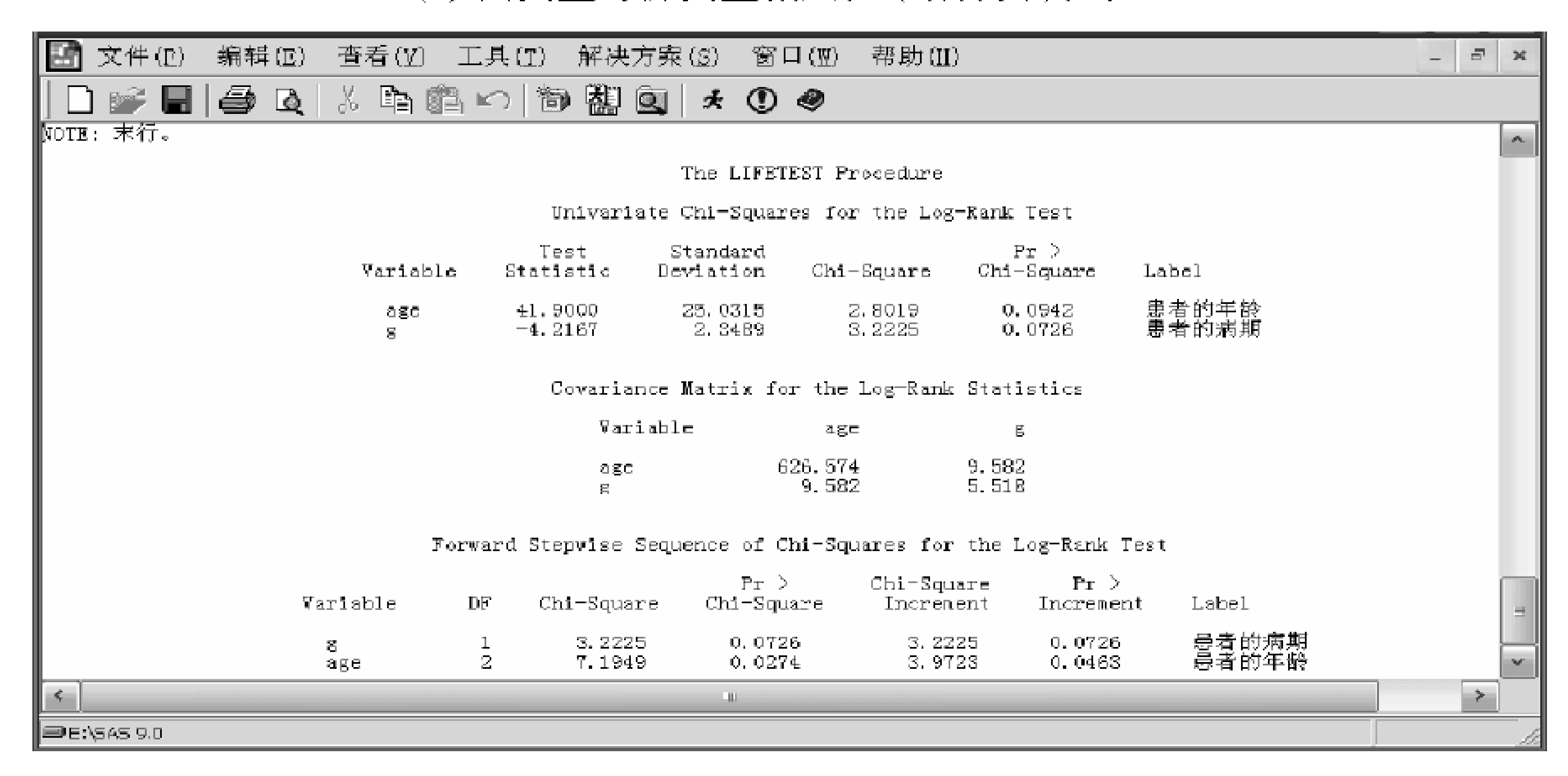
Covariance Matrix for the Wilcoxon Statistics

Variable	age	g
age	253.872	6.270
g	6.270	2.827

Forward Stepwise Sequence of Chi-Squares for the Wilcoxon Test

Variable	DF	Chi-Square	Pr > Chi-Square	Chi-Square Increment	Pr > Increment	Label
g	1	3.4946	0.0616	3.4946	0.0616	患者的病期
age	2	7.6411	0.0219	4.1465	0.0417	患者的年龄

(d) 因变量与协变量相关性(结合度)检验



NOTE: 末行。

The LIFETEST Procedure

Univariate Chi-Squares for the Log-Rank Test

Variable	Test Statistic	Standard Deviation	Chi-Square	Pr > Chi-Square	Label
age	41.9000	25.0315	2.8019	0.0942	患者的年龄
g	-4.2167	2.3489	3.2225	0.0726	患者的病期

Covariance Matrix for the Log-Rank Statistics

Variable	age	g
age	626.574	9.582
g	9.582	5.518

Forward Stepwise Sequence of Chi-Squares for the Log-Rank Test

Variable	DF	Chi-Square	Pr > Chi-Square	Chi-Square Increment	Pr > Increment	Label
g	1	3.2225	0.0726	3.2225	0.0726	患者的病期
age	2	7.1949	0.0274	3.9723	0.0463	患者的年龄

(e) 对数秩次卡方检验

图 13.8(续)

Univariate Chi-Squares for the Wilcoxon Test：对每个协变量，进行各层之间一致性的对数秩次卡方检验。

Covariance Matrix for the Wilcoxon Statistics：对数秩次的协方差矩阵。

Forward Stepwise Sequence of Chi-Squares for the Wilcoxon Test：逐步入选协变量时，对数秩次的卡方检验。从图形看，age 变量的“Pr>Chi-Square”值 0.0417 小于 α 值 0.05，说明年龄与手术后的寿命有关。

(4) 图 13.8(e)是因变量与协变量相关性(结合度)的 Log-Rank 检验，主要有以下几种检验：

Univariate Chi-Squares for the Log-Rank Test：对每个协变量，进行各层之间一致性的对数秩次卡方检验。

Covariance Matrix for the Log-Rank Statistics：对数秩次的协方差矩阵。

Forward Stepwise Sequence of Chi-Squares for the Log-Rank Test：逐步入选协变

量时，对数秩次的卡方检验。从图形看，age 变量的“Pr>Chi-Square”值 0.0463 小于 α 值 0.05，说明年龄与手术后的寿命有关。

例 10：肺癌数据的再分析。这里将 13.2.2 节中表 13.1 的肺癌数据，采用 LIFETEST 过程进行寿命检验，要求产生寿命分布函数、画出寿命曲线，并检验各个协变量与因变量的关系。

数据见 13.2.2 节中的例 6，LIFETEST 过程命令见下面 5 个语句。

```
PROC LIFETEST DATA=VALUNG2  OUTTEST=OUT3 PLOT=(S,LS);
    TIME t*CENSOR(1);
ID therapy;
STRATA cell;
      TEST cell kps diagtime age;
```

合并整理后新的程序命令见程序 13.6。

程序 13.6：

```
DATA valung2;
LABEL T='追踪时间或非追踪时间' kps='手术前的综合素质评分'
     diagtime='从诊断到手术治疗的时间'
     age='年龄' prior='事先是否有治疗' cell='细胞类型'
     therapy='疗法';
INPUT therapy cell t kps diagtime age prior $ @@ ;
  CENSOR=(t<0);
  t=ABS(t);
CARDS;

1 1  072 60  7 69 n 1 1  411 70 05 64 y 1 1  228 60  3 38 n
1 1  126 60  9 63 y 1 1  118 70 11 65 y 1 1   10 20  5 49 n
1 1  082 40 10 69 y 1 1  110 80 29 68 n 1 1  314 50 18 43 y
1 1 -100 70 06 70 n 1 1  042 60 04 81 n 1 1  008 40 58 63 y
1 1  144 30  4 63 n 1 1  -25 80  9 52 y 1 1   11 70 11 48 y
1 2   30 60  3 61 n 1 2  384 60  9 42 n 1 2   04 40 02 35 n
1 2   54 80  4 63 y 1 2   13 60  4 56 n 1 2 -123 40 03 55 n
1 2  -97 60  5 67 n 1 2  153 60 14 63 y 1 2   59 30  2 65 n
1 2  117 80  3 46 n 1 2  016 30 04 53 y 1 2  151 50 12 69 n
1 2   22 60  4 68 n 1 2   56 80 12 43 y 1 2   21 40  2 55 y
1 2   18 20 15 42 n 1 2  139 80 02 64 n 1 2   20 30  5 65 n
1 2   31 75  3 65 n 1 2  052 70 02 55 n 1 2  287 60 25 66 y
1 2   18 30  4 60 n 1 2   51 60  1 67 n 1 2  122 80 28 53 n
1 2   27 60  8 62 n 1 2   54 70  1 67 n 1 2  007 50  7 72 n
1 2   63 50 11 48 n 1 2  392 40 04 68 n 1 2   10 40 23 67 y
1 4   08 20 19 61 y 1 4   92 70 10 60 n 1 4   35 40  6 62 n
1 4  117 80 02 38 n 1 4  132 80  5 50 n 1 4   12 50  4 63 y
1 4  162 80  5 64 n 1 4  003 30 03 43 n 1 4   95 80  4 34 n
1 3  177 50 16 66 y 1 3  162 80  5 62 n 1 3  216 50 15 52 n
```

```
1  3   553  70   2  47 n 1 3   278  60  12  63 n 1 3   012  40  12  68 y
1  3   260  80   5  45 n 1 3   200  80  12  41 y 1 3   156  70   2  66 n
1  3  -182  90   2  62 n 1 3   143  90   8  60 n 1 3   105  80  11  66 n
1  3   103  80   5  38 n 1 3   250  70   8  53 y 1 3   100  60  13  37 y
2  1   999  90  12  54 y 2 1   112  80   6  60 n 2 1   -87  80   3  48 n
2  1  -231  50   8  52 y 2 1   242  50   1  70 n 2 1   991  70   7  50 y
2  1   111  70   3  62 n 2 1   001  20  21  65 y 2 1   587  60   3  58 n
2  1   389  90   2  62 n 2 1   033  30  06  64 n 2 1    25  20  36  63 n
2  1   357  70  13  58 n 2 1   467  90   2  64 n 2 1   201  80  28  52 y
2  1   001  50   7  35 n 2 1    30  70  11  63 n 2 1   044  60  13  70 y
2  1   283  90   2  51 n 2 1    15  50  13  40 y 2 3   049  30  03  37 n
2  2    25  30   2  69 n 2 2  -103  70  22  36 y 2 2    21  20  04  71 n
2  2    13  30   2  62 n 2 2   087  60  02  60 n 2 2    02  40  36  44 y
2  2    20  30   9  54 y 2 2   007  20  11  66 n 2 2    24  60   8  49 n
2  2    99  70   3  72 n 2 2   008  80  02  68 n 2 2    99  85   4  62 n
2  2    61  70   2  71 n 2 2   025  70  02  70 n 2 2    95  70   1  61 n
2  2    80  50  17  71 n 2 2   051  30  87  59 y 2 2    29  40   8  67 n
2  4    24  40  02  60 n 2 4   018  40  05  69 y 2 4   -83  99   3  57 n
2  4    31  80  03  39 n 2 4   051  60  05  62 n 2 4    90  60  22  50 y
2  4    52  60  03  43 n 2 4   073  60  03  70 n 2 4    08  50  05  66 n
2  4    36  70  08  61 n 2 4   048  10  04  81 n 2 4    07  40  04  58 n
2  4   140  70  03  63 n 2 4   186  90  03  60 n 2 4    84  80   4  62 n
2  4   019  50  10  42 n 2 4    45  40  03  69 n 2 4    80  40  04  63 n
2  3   052  60  04  45 n 2 3   164  70  15  68 y 2 3    19  30  04  39 y
2  3   053  60  12  66 n 2 3   015  30  05  63 n 2 3    43  60  11  49 y
2  3   340  80  10  64 y 2 3   133  75  01  65 n 2 3   111  60  05  64 n
2  3   231  70  18  67 y 2 3   378  80  04  65 n
;
PROC LIFETEST DATA=VALUNG2 OUTTEST=OUT3 PLOT=(S,LS);
    TIME t*CENSOR(1);
ID therapy;
STRATA prior;
TEST cell kps;      /* therapy diagtime age prior 等变量与寿命无关而删除 */
RUN;
```

运行程序 13.6 产生图 13.9、图 13.10 所示的结果。

对图 13.9 的结果分析：

如图 13.9(b)所示：

(1) Stratum 2：prior = y

本例分为“手术之前有治疗过”与“没有治疗”两层数据，然后进行寿命检验。

(2) Product-Limit Survival Estimates

采用极限乘法寿命估计。其中：

T 因变量　寿命时间

Survival　寿命概率

Failure 死亡概率

Survival Standard Error 寿命标准误差

Number Failed 死亡数

Number Left 生存数

Therapy 疗法,1 为标准疗法,2 为试验疗法

文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)

① Stratum 2: prior = y

②Product-Limit Survival Estimates

T	Survival	Failure	Survival Standard Error	Number Failed	Number Left	Therapy
0.000	1.0000	0	0	0	40	.
1.000	0.9750	0.0250	0.0247	1	39	2
2.000	0.9500	0.0500	0.0345	2	38	2
8.000	.	.	.	3	37	1
8.000	0.9000	0.1000	0.0474	4	36	1
10.000	0.8750	0.1250	0.0523	5	35	1
11.000	0.8500	0.1500	0.0565	6	34	1
12.000	.	.	.	7	33	1
......						
21.000	0.6500	0.3500	0.0754	14	26	1
25.000*	.	.	.	14	25	1
43.000	0.6240	0.3760	0.0767	15	24	2
44.000	0.5980	0.4020	0.0778	16	23	2
51.000	0.5720	0.4280	0.0787	17	22	2
54.000	0.5460	0.4540	0.0793	18	21	1
56.000	0.5200	0.4800	0.0796	19	20	1
82.000	0.4940	0.5060	0.0798	20	19	1
90.000	0.4680	0.5320	0.0797	21	18	2
100.000	0.4420	0.5580	0.0794	22	17	1
103.000*	.	.	.	22	16	2
118.000	0.4144	0.5856	0.0791	23	15	1
126.000	0.3867	0.6133	0.0785	24	14	1
153.000	0.3591	0.6409	0.0776	25	13	1
164.000	0.3315	0.6685	0.0764	26	12	2
177.000	0.3039	0.6961	0.0749	27	11	1
200.000	0.2762	0.7238	0.0730	28	10	1
201.000	0.2486	0.7514	0.0707	29	9	2
231.000	0.2210	0.7790	0.0680	30	8	2
231.000*	.	.	.	30	7	2
250.000	0.1894	0.8106	0.0652	31	6	1
287.000	0.1579	0.8421	0.0615	32	5	1
314.000	0.1263	0.8737	0.0567	33	4	1
340.000	0.0947	0.9053	0.0506	34	3	2
411.000	0.0631	0.9369	0.0424	35	2	1
991.000	0.0316	0.9684	0.0308	36	1	2
999.000	0	1.0000	0	37	0	2

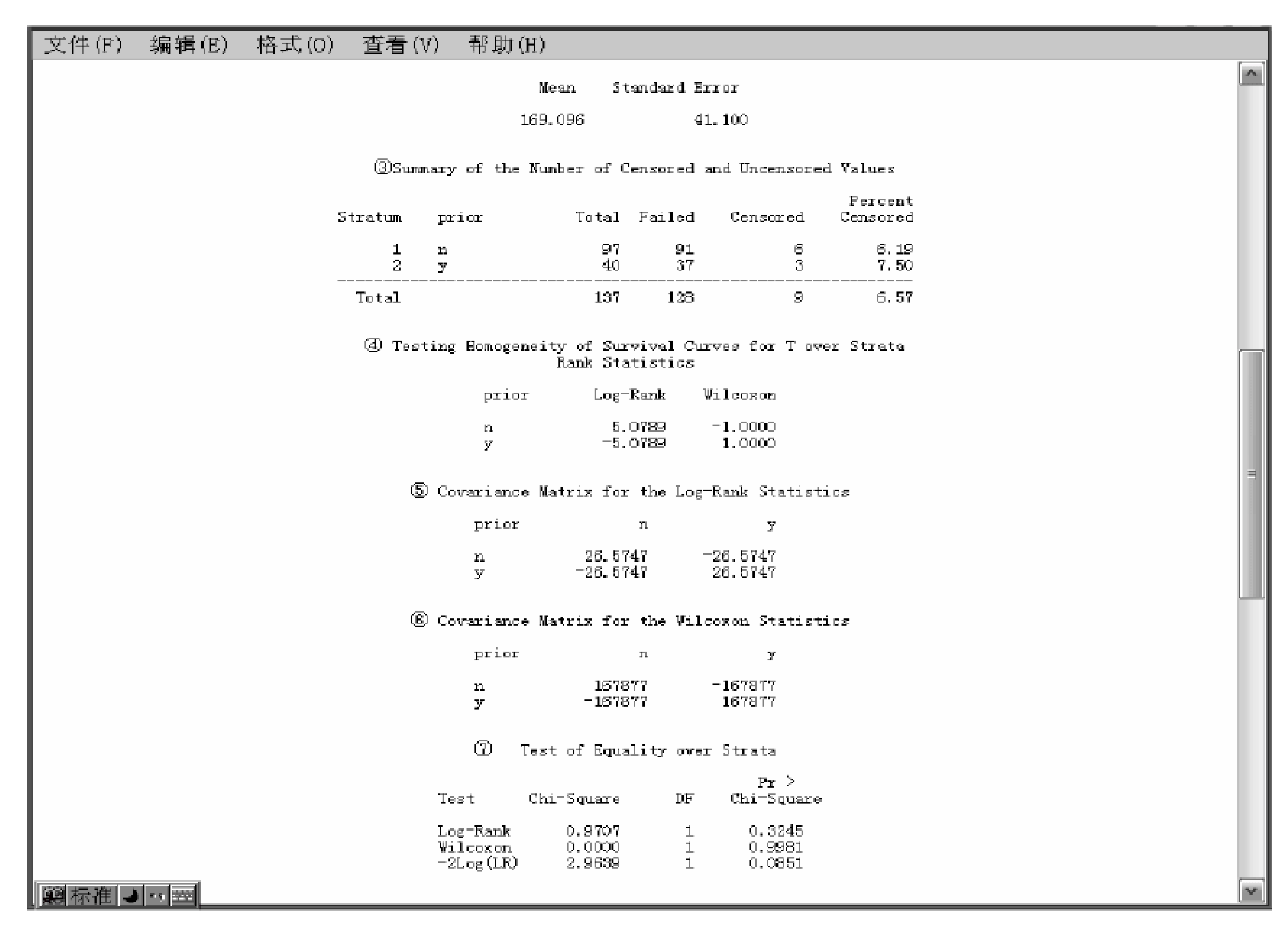
NOTE: The marked survival times are censored observations.

Summary Statistics for Time Variable T

Quartile Estimates

Percent	Point Estimate	95% Confidence Interval [Lower	Upper)
75	201.000	126.000	314.000
50	82.000	43.000	164.000
25	17.000	11.000	51.000

(a) 极限乘法寿命分布



文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)

Mean	Standard Error
169.096	41.100

③Summary of the Number of Censored and Uncensored Values

Stratum	prior	Total	Failed	Censored	Percent Censored
1	n	97	91	6	6.19
2	y	40	37	3	7.50
Total		137	128	9	6.57

④ Testing Homogeneity of Survival Curves for T over Strata

Rank Statistics

prior	Log-Rank	Wilcoxon
n	5.0789	-1.0000
y	-5.0789	1.0000

⑤ Covariance Matrix for the Log-Rank Statistics

prior	n	y
n	26.5747	-26.5747
y	-26.5747	26.5747

⑥ Covariance Matrix for the Wilcoxon Statistics

prior	n	y
n	157877	-167877
y	-157877	167877

⑦ Test of Equality over Strata

Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	0.9707	1	0.3245
Wilcoxon	0.0000	1	0.9981
-2Log(LR)	2.9639	1	0.0851

(b) 各层间一致性的 3 种检验法

图 13.9 肺癌数据的分层寿命检验

```
文件(F)  编辑(E)  格式(O)  查看(V)  帮助(H)

              ⑧ Rank Tests for the Association of T with Covariates Pooled over Strata
                          Univariate Chi-Squares for the Wilcoxon Test

                    Test       Standard                  Pr >
     Variable    Statistic    Deviation   Chi-Square   Chi-Square    Label
       cell       -8.5243       6.9667      1.4971       0.2211      细胞类型
       kps        919.8         119.5      59.2303      <.0001       手术前的综合素质评分

                         Covariance Matrix for the Wilcoxon Statistics
                           Variable            cell          kps

                           cell                48.5          35.6
                           kps                 35.6       14282.7
                ⑨ Forward Stepwise Sequence of Chi-Squares for the Wilcoxon Test

                                          Pr >       Chi-Square     Pr >
  Variable      DF    Chi-Square   Chi-Square    Increment   Increment    Label

    kps          1      59.2303      <.0001       59.2303      <.0001     手术前的综合素质评分
    cell         2      61.6463      <.0001        2.4160      0.1201     细胞类型

                                        SAS 系统                10:04 Saturday, May 4, 2002  57

                                 The LIFETEST Procedure

                         Univariate Chi-Squares for the Log-Rank Test

                    Test       Standard                  Pr >
     Variable    Statistic    Deviation   Chi-Square   Chi-Square    Label

       cell      -25.4939      11.3248      5.0678       0.0244      细胞类型
       kps       1200.4        182.7       43.1727      <.0001       手术前的综合素质评分

                        Covariance Matrix for the Log-Rank Statistics

                           Variable            cell          kps

                           cell               128.3         152.1
                           kps                152.1       33375.2

                ⑩ Forward Stepwise Sequence of Chi-Squares for the Log-Rank Test

                                          Pr >       Chi-Square     Pr >
  Variable      DF    Chi-Square   Chi-Square    Increment   Increment    Label

    kps          1      43.1727      <.0001       43.1727      <.0001     手术前的综合素质评分
    cell         2      50.6899      <.0001        7.5172      0.0061     细胞类型
```

(c) 与因变量有关的协变量

图 13.9（续）

(3) Summary of the Number of Censored and Uncensored Values

各层追踪人数占层内人数的比例。如图 13.9(b)所示，第二层(预先有治疗)有 7.50%，第一层(预先没有治疗)有 6.19%。

(4) Testing Homogeneity of Survival Curves for T over Strata

各层寿命函数一致性的检验。主要结果包括以下(5)～(7)项。

(5) Covariance Matrix for the Log-Rank Statistics

对数秩次的协方差矩阵。

(6) Covariance Matrix for the Wilcoxon Statistics

基于 Wilcoxon 的协方差矩阵。

(7) Test of Equality over Strata

各层寿命函数一致性的检验。由“Pr> Chi-Square”一栏的 3 个值看出，它们分别大于 α 值 0.05，所以不能拒绝“各层寿命一致性”的假设。

(8) Rank Tests for the Association of T with Covariates Pooled over Strata

每层中，协变量与因变量(寿命)相关性的秩次检验，主要统计量是：

Univariate Chi-Squares for the Wilcoxon Test：单变量 Wilcoxon 卡方检验。只有 kps 变量的“Pr> Chi-Square”值显著(<.0001)，表明与寿命时间有关。

(9) Forward Stepwise Sequence of Chi-Squares for the Wilcoxon Test

逐步入选变量时的 Wilcoxon 卡方检验。从“Pr> Chi-Square”值(<.0001)看，也只有 kps 变量与寿命时间有关。

Univariate Chi-Squares for the Log-Rank Test：单变量 Log-Rank(对数秩次)卡方检验。Cell 变量和 kps 变量的“Pr> Chi-Square”值都很显著(小于 α 值 0.05)，表明这两

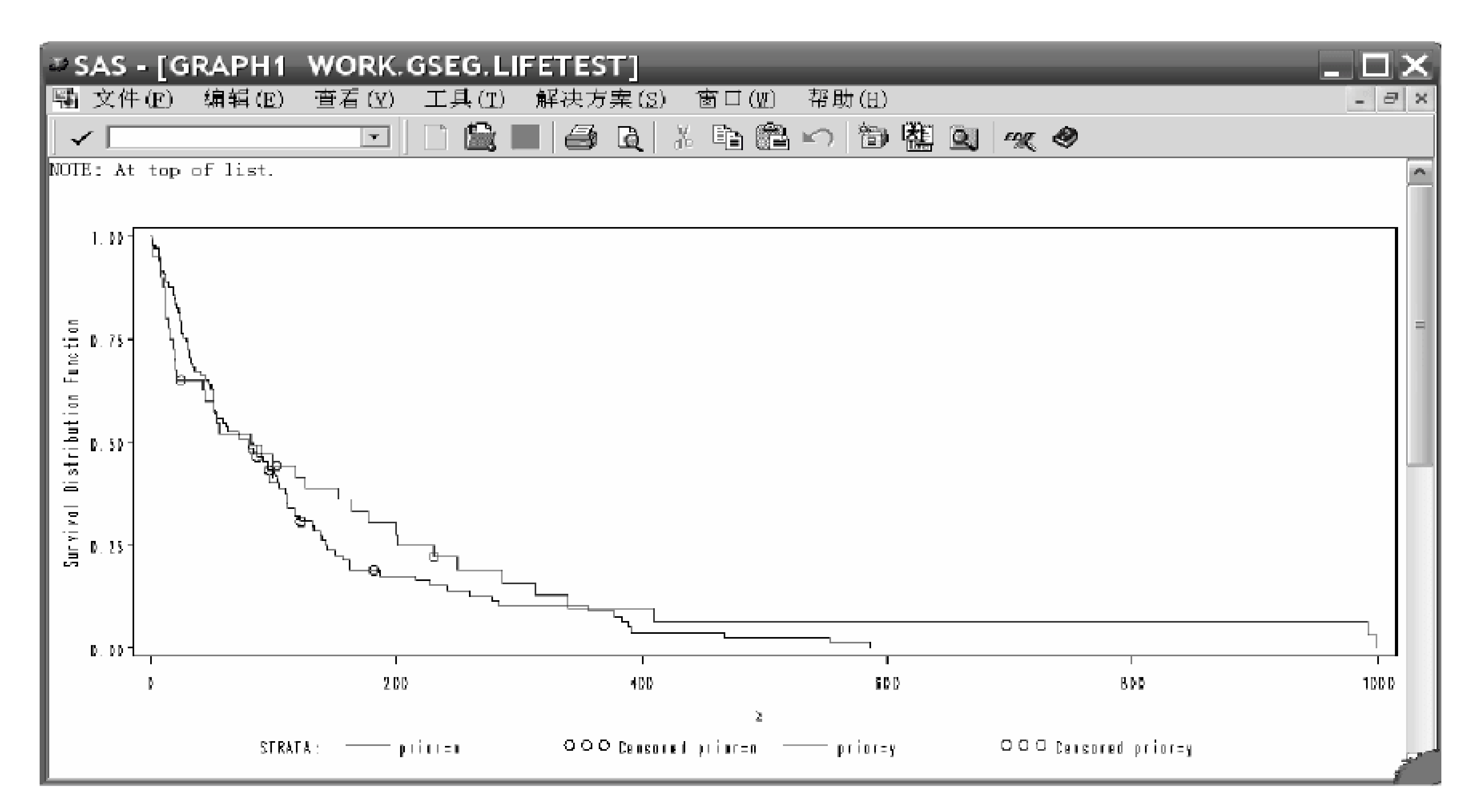
个变量与寿命时间都有关。

说明：相比之下，(9)项比(8)项宽容。

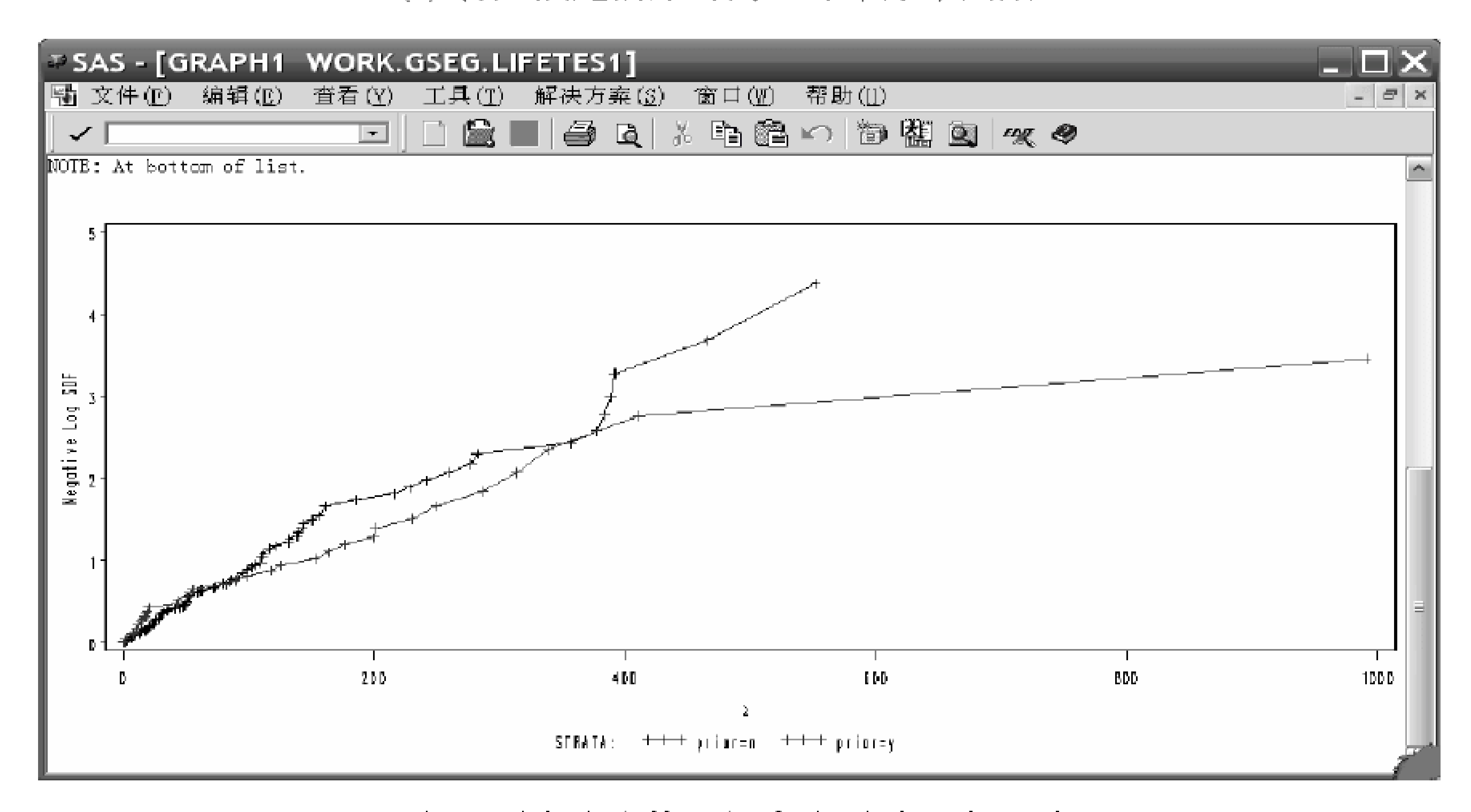
(10) Forward Stepwise Sequence of Chi-Squares for the Log-Rank Test

逐步入选变量时的对数秩次(Log-Rank)检验。Cell 变量和 kps 变量的“Pr> Chi-Square”值都很显著(小于 α 值 0.05)，表明这两个变量与寿命时间都有关。

比较(8)～(10)项可以看出，Log-Rank(对数秩次)寿命检验，总是比 Wilcoxon 寿命检验宽容。



(a) 有无预先治疗的两组寿命分布函数



(b)有无预先治疗的两组寿命对数分布函数

图 13.10 分两层的寿命分布函数

对图 13.10 的结果分析：

图 13.10(a)是寿命分布曲线。从图 13.10(a)看，手术前预先做过辅助治疗一组的肺癌患者(prior=yes)，其寿命分布曲线比较平坦，待存活 400 天后基本上呈现平稳状态，且寿命时间长达 900 多天。但手术前未做辅助治疗一组的肺癌患者(prior=no)，其寿命

分布曲线比较陡峭，到存活 400 天后虽然也趋于平稳，但寿命时间最长只有 500 多天。

图 13.10(b)是自然对数的寿命分布曲线。从图 13.10(b)看，手术前预先做过辅助治疗一组的肺癌患者(prior＝yes)，其寿命分布曲线比较平坦，且寿命时间长达 900 多天。但手术前未做辅助治疗一组的肺癌患者(prior＝no)，其寿命分布曲线比较陡峭而短促，寿命时间最长只有 500 多天。

习 题 13

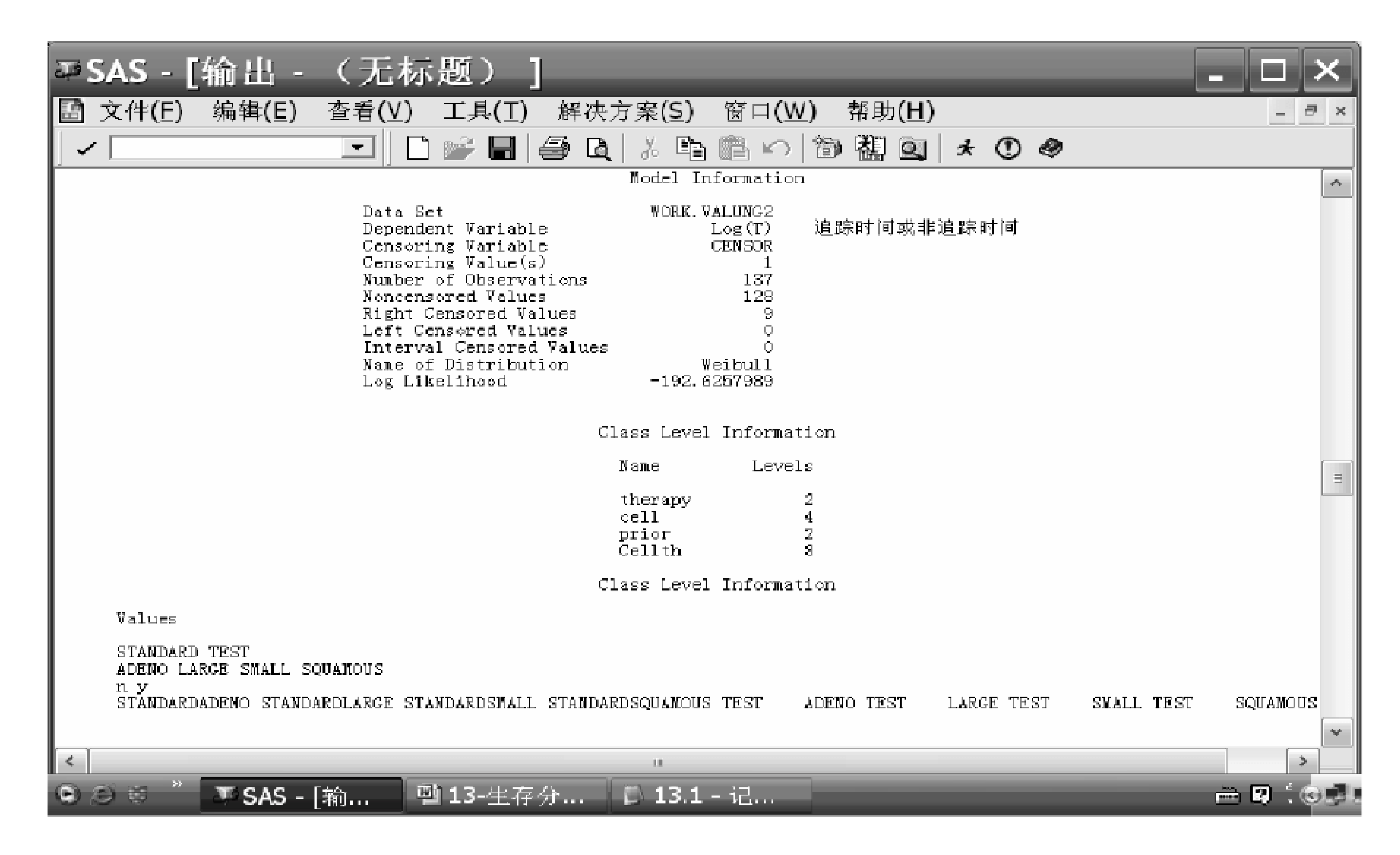
1. 下面程序 13.7 是对程序 13.2(肺癌数据)的另一种解法。

程序 13.7：

```
DATA valung2;
DROP I N;
INPUT therapy $  cell $  n @  ;
  Cellth= therapy ||cell;
  LABEL T= '追踪时间或非追踪时间' kps= '手术前的综合素质评分'
  diagtime= '从诊断到手术治疗的时间'
       age= '年龄' prior= '事先是否有治疗' cell= '细胞类型'
       therapy= '疗法';
DO I=1 TO N;
INPUT t kps diagtime age prior $  @@ ;
  CENSOR= (t< 0);
  t= ABS(t);
OUTPUT;
END;
CARDS;
STANDARD SQUAMOUS 15
  072   60    7   69   n    411   70   05   64   y   228   60    3   38   n
  126   60    9   63   y    118   70   11   65   y    10   20    5   49   n
  082   40   10   69   y    110   80   29   68   n   314   50   18   43   n
- 100   70   06   70   n    042   60   04   81   n   008   40   58   63   y
  144   30   04   63   n   - 25   80    9   52   y    11   70   11   48   y
STANDARD SMALL 30
 030   60    3   61   n   384   60    9   42   n     04   40   02   35   n
  54   80    4   63   y    13   60    4   56   n  - 123   40   03   55   n
- 97   60    5   67   n   153   60   14   63   y     59   30    2   65   n
 117   80    3   46   n   016   30   04   53   y    151   50   12   69   n
  22   60    4   68   n    56   80   12   43   y     21   40    2   55   y
  18   20   15   42   n   139   80   02   64   n     20   30    5   65   n
  31   75    3   65   n   052   70   02   55   n    287   60   25   66   y
  18   30    4   60   n    51   60    1   67   n    122   80   28   53   n
  27   60    8   62   n    54   70    1   67   n    007   50    7   72   n
  63   50   11   48   n   392   40   04   68   n     10   40   23   67   y
STANDARD ADENO 9
```

```
 08   20   19   61   y    92   70   10   60   n   35   40   6   62   n
117   80   02   38   n   132   80    5   50   n   12   50   4   63   y
162   80    5   64   n   003   30   03   43   n   95   80   4   34   n
STANDARD LARGE 15
  177   50   16   66   y   162   80    5   62   n   216   50   15   52   n
  553   70    2   47   n   278   60   12   63   n   012   40   12   68   y
  260   80    5   45   n   200   80   12   41   y   156   70    2   66   n
-182    90    2   62   n   143   90    8   60   n   105   80   11   66   n
  103   80    5   38   n   250   70    8   53   y   100   60   13   37   y
TEST SQUAMOUS 20
  999   90   12   54   y   112   80    6   60   n   -87   80    3   48   n
-231    50    8   52   y   242   50    1   70   n   991   70    7   50   y
  111   70    3   62   n   001   20   21   65   y   587   60    3   58   n
  389   90    2   62   n   033   30   06   64   n    25   20   36   63   n
  357   70   13   58   n   467   90    2   64   n   201   80   28   52   y
  001   50    7   35   n    30   70   11   63   n    44   60   13   70   y
  283   90    2   51   n    15   50   13   40   y
TEST SMALL 18
 25  30    2   69   n   -103   70   22   36   y   21   20   04   71   n
 13  30    2   62   n    087   60   02   60   n   02   40   36   44   y
 20  30    9   54   y    007   20   11   66   n   24   60    8   49   n
 99  70    3   72   n    008   80   02   68   n   99   85    4   62   n
 61  70    2   71   n    025   70   02   70   n   95   70    1   61   n
 80  50   17   71   n    051   30   87   59   y   29   40    8   67   n
TEST ADENO 18
 24   40   02   60   n   018   40   05   69   y   -83   99    3   57   n
 31   80   03   39   n   051   60   05   62   n    90   60   22   50   y
 52   60   03   43   n   073   60   03   70   n    08   50   05   66   n
 36   70   08   61   n   048   10   04   81   n    07   40   04   58   n
140   70   03   63   n   186   90   03   60   n    84   80    4   62   n
019   50   10   42   n    45   40   03   69   n    80   40   04   63   n
TEST LARGE 12
052   60   04   45   n   164   70   15   68   y    19   30   04   39   y
053   60   12   66   n   015   30   05   63   n    43   60   11   49   y
340   80   10   64   y   133   75   01   65   n   111   60   05   64   n
231   70   18   67   y   378   80   04   65   n   049   30   03   37   n
 ;
PROC FORMAT;
   VALUE CELLf 1='鱼鳞状'  2='小型'  3='大型'  4='腺状';
   FORMAT CELL CELLf.;
PROC LIFEREG;
      CLASS therapy cell prior cellth;
   MODEL t*CENSOR(1)=therapy cell prior kps age diagtime cellth/D=WEIBULL;
OUTPUT OUT=OUT2 P=PRED;
PROC PRINT;
RUN;
```

运行程序 13.7 产生图 13.11 所示的结果。



Model Information

Data Set	WORK.VALUNG2	
Dependent Variable	Log(T)	追踪时间或非追踪时间
Censoring Variable	CENSOR	
Censoring Value(s)	1	
Number of Observations	137	
Noncensored Values	128	
Right Censored Values	9	
Left Censored Values	0	
Interval Censored Values	0	
Name of Distribution	Weibull	
Log Likelihood	-192.6257989	

Class Level Information

Name	Levels
therapy	2
cell	4
prior	2
Cellth	8

Class Level Information

Values

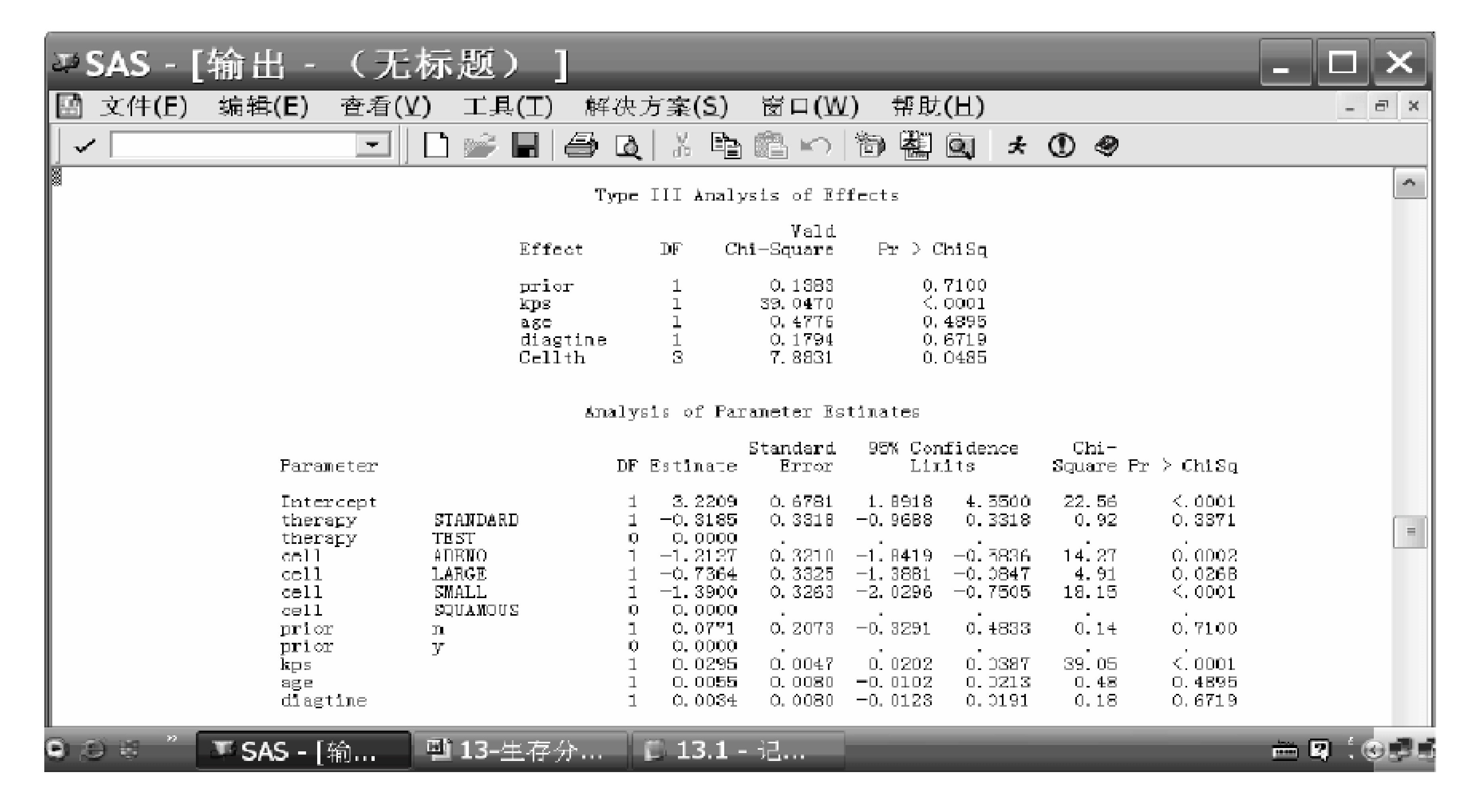
STANDARD TEST
ADENO LARGE SMALL SQUAMOUS
n y
STANDARDADENO STANDARDLARGE STANDARDSMALL STANDARDSQUAMOUS TEST ADENO TEST LARGE TEST SMALL TEST SQUAMOUS

(a) Weibull模型采用极值基准分布拟合



Parameter			DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Cellth	STANDARDADENO		1	0.1146	0.5145	-0.8938	1.1231	0.05	0.8237
Cellth	STANDARDLARGE		1	0.7783	0.4825	0.1673	1.7239	2.60	0.1067
Cellth	STANDARDSMALL		1	1.0586	0.4376	0.2009	1.9163	5.85	0.0156
Cellth	STANDARDSQUAMOUS		0	0.0000	.	.	.	.	.
Cellth	TEST	ADENO	0	0.0000	.	.	.	.	.
Cellth	TEST	LARGE	0	0.0000	.	.	.	.	.
Cellth	TEST	SMALL	0	0.0000	.	.	.	.	.
Cellth	TEST	SQUAMOUS	0	0.0000	.	.	.	.	.
Scale			1	0.8861	0.0607	0.7748	1.0133		
Weibull Shape			1	1.1286	0.0773	0.9869	1.2906		

(b) 各单元搭配后的显著性检验



Type III Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
prior	1	0.1383	0.7100
kps	1	39.0470	<.0001
age	1	0.4776	0.4895
diagtime	1	0.1794	0.6719
Cellth	3	7.8831	0.0485

Analysis of Parameter Estimates

Parameter		DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept		1	3.2209	0.6781	1.8918	4.5500	22.56	<.0001
therapy	STANDARD	1	-0.3185	0.3318	-0.9688	0.3318	0.92	0.3371
therapy	TEST	0	0.0000	.	.	.	.	.
cell	ADENO	1	-1.2127	0.3210	-1.8419	-0.5836	14.27	0.0002
cell	LARGE	1	-0.7364	0.3325	-1.3881	-0.0847	4.91	0.0268
cell	SMALL	1	-1.3900	0.3263	-2.0296	-0.7505	18.15	<.0001
cell	SQUAMOUS	0	0.0000	.	.	.	.	.
prior	n	1	0.0771	0.2073	-0.3291	0.4833	0.14	0.7100
prior	y	0	0.0000	.	.	.	.	.
kps		1	0.0295	0.0047	0.0202	0.0387	39.05	<.0001
age		1	0.0055	0.0080	-0.0102	0.0213	0.48	0.4895
diagtime		1	0.0034	0.0080	-0.0123	0.0191	0.18	0.6719

(c) Weibull模型参数估计

图 13.11 肺癌患者寿命分析

试对产生的结果进行比较。

2. 表 13.2 中的数据是某医院糖尿病患者的随访情况。

表 13.2 某医院糖尿病患者随访情况

发病年数	存活人数	死亡人数	病号
15	146	2	1
16	144	4	2
17	140	8	3
18	136	10	4
19	130	12	5
20	121	15	6
21	115	20	7
22	101	24	8
23	91	30	9
24	80	46	10

试做寿命估计并计算寿命曲线。

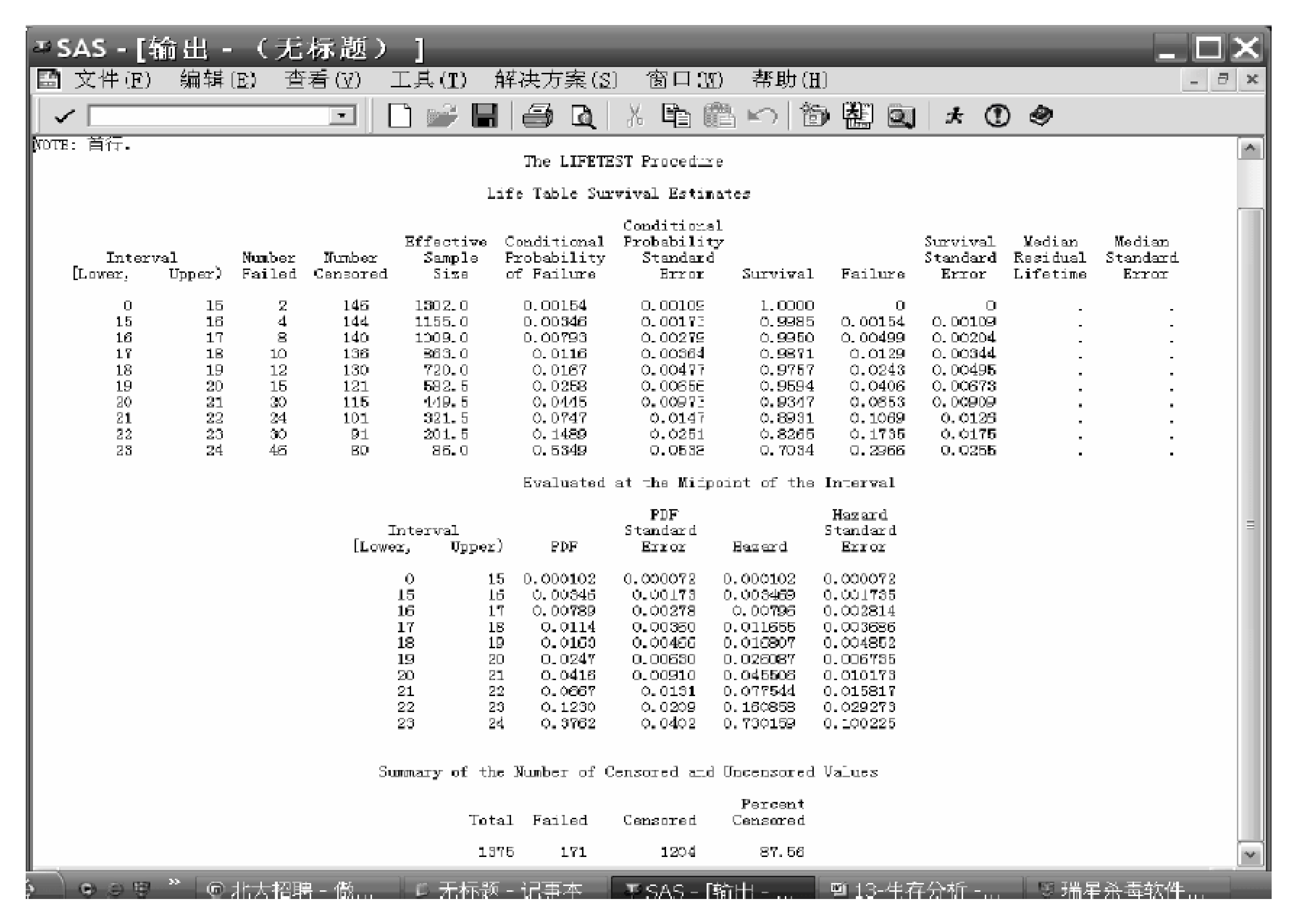
根据表 13.2 中的数据及题意,编出过程命令见程序 13.8。

程序 13.8:

```
DATA T;
LABEL year='发病年数'  v1='存活人数'  v2='死亡人数';
  DO year=14.5 TO 23.5;            /* 时间范围为 15 年~24 年,取中点值则为 14.5 TO 23.5,步长默
                                      认为 1 年 */
    INPUT v1 v2@ @ ;               /* 依次读取存活人数 v1 和死亡人数 v2 */
      Censor=1;count=v1;           /* 将存活人数赋予 v1 变量,并用 Censor= 1 记为
                                      追踪值 */
        OUTPUT;                    /* 输出结果 */
      Censor=2;count=v2;           /* 将死亡人数赋予 v2 变量,并用 Censor= 2 记为
                                      非追踪值 */
      OUTPUT;                      /* 输出结果 */
    DROP v1 v2;
END;
CARDS;
146 2 144 4 140 8 136 10 130 12
121 15 115 20 101 24 91 30 80 46
;
PROC LIFETEST PLOTS=(S,H)
        INTERVAL=(15 TO 24) METHOD=LT;      /* 指定区间范围和寿命表 */
TIME year* censor(1);              /* 指明 censor 为因变量, censor 为指示变量,
                                      当 censor=1 时为追踪值 */
    FREQ count;                    /* 对 count 变量进行频数统计 */
RUN;
```

运行程序 13.8 产生图 13.12 所示的结果。

请分析寿命表。



The LIFETEST Procedure

Life Table Survival Estimates

Interval [Lower,	Upper)	Number Failed	Number Censored	Effective Sample Size	Conditional Probability of Failure	Conditional Probability Standard Error	Survival	Failure	Survival Standard Error	Median Residual Lifetime	Median Standard Error
0	15	2	146	1302.0	0.00154	0.00109	1.0000	0	0	.	.
15	16	4	144	1155.0	0.00346	0.00173	0.9985	0.00154	0.00109	.	.
16	17	8	140	1309.0	0.00793	0.00275	0.9950	0.00499	0.00204	.	.
17	18	10	136	863.0	0.0116	0.00364	0.9871	0.0129	0.00344	.	.
18	19	12	130	720.0	0.0167	0.00477	0.9757	0.0243	0.00495	.	.
19	20	15	121	582.5	0.0258	0.00656	0.9594	0.0406	0.00673	.	.
20	21	30	115	419.5	0.0445	0.00973	0.9347	0.0653	0.00909	.	.
21	22	24	101	321.5	0.0747	0.0147	0.8931	0.1069	0.0125	.	.
22	23	30	91	201.5	0.1489	0.0251	0.8265	0.1735	0.0175	.	.
23	24	46	80	86.0	0.5349	0.0532	0.7034	0.2966	0.0255	.	.

Evaluated at the Midpoint of the Interval

Interval [Lower,	Upper)	PDF	PDF Standard Error	Hazard	Hazard Standard Error
0	15	0.000102	0.000072	0.000102	0.000072
15	16	0.00346	0.00173	0.003469	0.001735
16	17	0.00789	0.00278	0.00796	0.002814
17	18	0.0114	0.00360	0.011655	0.003686
18	19	0.0163	0.00466	0.016807	0.004852
19	20	0.0247	0.00630	0.026087	0.006735
20	21	0.0416	0.00910	0.045506	0.010173
21	22	0.0667	0.0131	0.077544	0.015817
22	23	0.1230	0.0209	0.160858	0.029273
23	24	0.3762	0.0402	0.730159	0.100225

Summary of the Number of Censored and Uncensored Values

Total	Failed	Censored	Percent Censored
1375	171	1204	87.56

图 13.12 寿命表

第 14 章 chapter 14

非线性回归分析一：对数与多项式回归

在计算机 SAS 中，非线性回归过程称为 PROC NLIN。其中的 NLIN 全称为 Non Linear Regression。非线性回归是通过复杂的迭代法获得的回归模型。

NLIN 过程的迭代法有以下 5 种。

（1）Gauss 法：高斯法。又称改良的高斯—牛顿法（Modified Gauss-Newton method）。

（2）Dud 法：错位法（False Position）和多元正割法（Multivariate Secant）。

（3）Gradient 法：梯度法（Gradient method），又称急剧下降法（Steepest Descent method）。

（4）Newton 法：牛顿法（Newton method）。

（5）Marquardt 法。

本章介绍的曲线回归包括对数回归和多项式回归。

14.1 对数曲线回归

对数曲线回归必须满足关系式：$Y=a+b\times\log(X)$。

14.1.1 对数曲线回归所要求的数据

下面对“产量与劳动力（变量 L）、产量与资金（变量 K）”的数据进行非线性回归。

（1）非线性回归的数据及程序见程序 14.1。

程序 14.1：产量与劳动力（L）、产量与资金（K）的数据与程序。

```
Title  '非线性回归 (NonLinear Regression)';
DATA zj;
INPUT k L LOGQ;
CARDS;

.228     .802    -1.359
.258     .249    -1.695
.767     .511     -.649
.495     .758     -.165
```

```
.487      .425      -.270
.678      .452      -.473
.784      .817       .031
.727      .845      -.563
.695      .958      -.125
.458      .084      -2.218
.981      .021      -3.633
.002      .295      -5.586
.429      .277      -0.773
.231      .546      -1.315
.644      .129      -1.678
.631      .017      -3.879
.059      .906      -2.301
.811      .223      -1.377
.758      .145      -2.27
.050      .161      -2.539
.823      .006      -5.15
.483      .836      -.324
.682      .521      -.253
.116      .930      -1.53
.440      .495      -.614
.456      .185      -1.151
.342      .092      -2.089
.358      .485      -.951
.162      .934      -1.275
;
PROC NLIN BEST=100 METHOD=DUD;
MODEL LOGQ=B0+C*LOG(D*(L**r)+(1-D)*(K**r));
PARMS B0=1 C=-1 D=.5 R=-1;
RUN;
```

(2) 从程序 14.1 中的数据发掘非线性回归模型如下：

$$\mathrm{LOGQ} = B_0 + C \times \log(D \times (L^r) + (1 - D) \times (K^r)) \tag{14.1}$$

(3) 参数说明：

B_0：截距。 D：分布参数。

C：斜率，即效率参数。 r：替代参数。

(4) 对本例的产量与劳动力(变量 L)、产量与资金(变量 K)的对数回归分析见以下各节。

14.1.2 对数曲线回归的编程解法

操作步骤如下：

(1) 将程序 14.1 中的语句与数据调入 SAS 的程序编辑器(Editor Program)窗口。

(2) 选择“运行”→“提交”命令，SAS 输出图 14.1～图 14.5 所示的结果。

14.2　对数曲线回归分析

下面对图 14.1～图 14.5 所示的结果进行分析。

非线性回归(NonLinear Regression)
09:05 Saturday, June 15, 2002

The NLIN Procedure
Dependent Variable LOGq
Grid Search

B0	C	D	r	Sum of Squares
1.0000	－1.0000	0.5000	－1.0000	36.7867

图 14.1　因变量与模型的系数

非线性回归(NonLinear Regression)
The NLIN Procedure
Dependent Variable LOGq
DUD Initialization

DUD	B0	C	D	r	Sum of Squares
－5	1.0000	－1.0000	0.5000	－1.0000	36.7867
－4	1.1000	－1.0000	0.5000	－1.0000	43.3528
－3	1.0000	－1.1000	0.5000	－1.0000	27.4043
－2	1.0000	－1.0000	0.5500	－1.0000	36.2204
－1	1.0000	－1.0000	0.5000	－1.1000	26.3184

图 14.2　虚点系数的初始值

图 14.1 和图 14.2 显示出模型的初始参数值。

从图 14.3 的统计量看：迭代回归了 13 步，因为最小平方和已经不能达到收敛标准，不能再下降了，因此迭代终止。

对图 14.4(b)中统计量的解释如下。

(1) 方差分析(以图 14.4(b)为准)

Model Sum of Squares (Regression SS)：是已被解释的回归平方和(SAS 9e 输出的是 56.5597)。

Error Sum of Squares(或显示 Residual Sum of Squares)：残差平方和。

Residual Sum of Squares (即 Residual SS)：未被解释的残差平方和(其值为 1.6492)。

Uncorrected Total Sum of Squares：因变量的总平方和(此项见图 14.4(a)，其值为 131.7)。



非线性回归(NonLinear Regression) 14:15 Monday, January

The NLIN Procedure
Dependent Variable LOGQ
Method: Gauss-Newton

Iterative Phase

Iter	B0	C	D	r	Sum of Squares
0	1.0000	-1.0000	0.5000	-1.0000	36.7867
1	0.7593	-0.7322	0.5248	-1.2549	27.9946
2	0.5207	-0.4963	0.5627	-1.6784	21.6044
3	0.1872	-0.2760	0.6495	-2.6578	14.8316
4	0.0983	-0.3607	0.6655	-2.7985	1.6498
5	0.0950	-0.3709	0.6597	-2.7204	1.6493
6	0.0941	-0.3748	0.6576	-2.6936	1.6492
7	0.0939	-0.3735	0.6582	-2.7024	1.6492
8	0.0939	-0.3741	0.6579	-2.6985	1.6492
9	0.0939	-0.3738	0.6581	-2.7005	1.6492
10	0.0939	-0.3740	0.6580	-2.6995	1.6492
11	0.0939	-0.3739	0.6581	-2.7000	1.6492
12	0.0939	-0.3739	0.6580	-2.6998	1.6492
13	0.0939	-0.3739	0.6580	-2.6999	1.6492

图 14.3　各个迭代阶段的系数值

Corrected Total Sum of Squares：偏离均值的平方和(其值为 58.2089)。

R squared＝1－Residual SS/Corrected SS：确定系数，或称判定系数为 0.97。此值表示被模型解释的方差占总方差的 97%(图 14.4(b)中不显示此项)。

本例中 R squared＝0.97 接近于 1，说明模型很好地拟合了数据。

(2) 置信区间分析

Estimation Summary

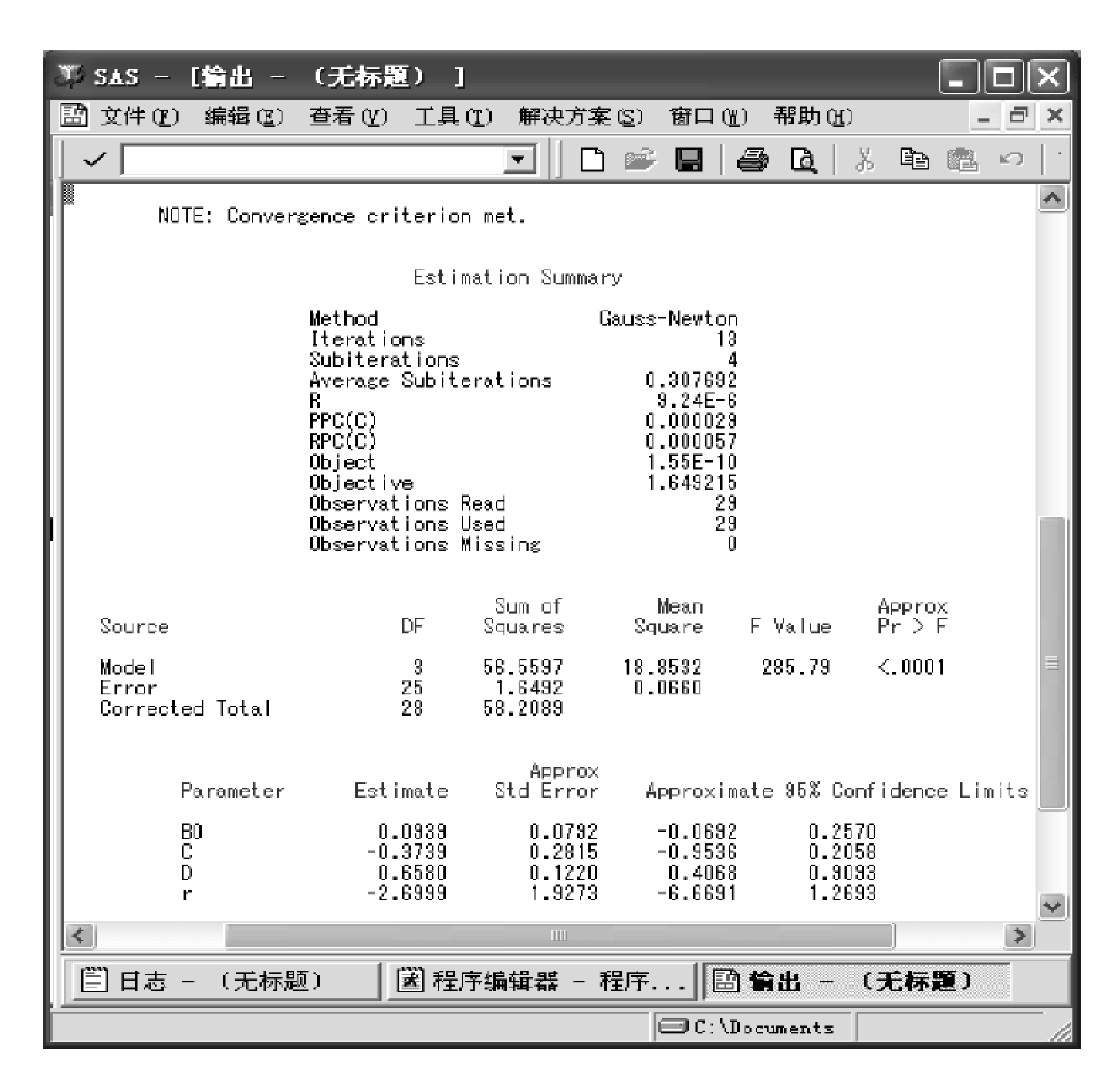
Method	DUD
Iterations	37
Object	1.502E-9
Objective	1.649215
Observations Read	29
Observations Used	29
Observations Missing	0

Source	DF	Sum of Squares	Mean Square	F Value	Approx Pr > F
Regression	4	130.1	32.5196	285.79	<.0001
Residual	25	1.6492	0.0660		
Uncorrected Total	29	131.7			
Corrected Total	28	58.2089			

Parameter	Estimate	Approx Std Error	Approximate 95% Confidence Limits	
B0	0.0939	0.0792	-0.0692	0.2570
C	-0.3739	0.2815	-0.9535	0.2058
D	0.6580	0.1220	0.4068	0.9093
r	-2.6999	1.9272	-6.6691	1.2692

(a) SAS 9 以前版本产生的回归输出

图 14.4　回归平方和



(b) SAS 9版本产生的回归输出

图 14.4（续）

从图 14.4(b)底部的统计量看：

Parameter：参数；有 B_0、C、D、r 四个参数；

Estimate：参数的估计值，如 B_0 参数的估计值为 0.0939。C 参数的估计值为 −0.3739 等。

Approx Std Error：逼近的标准误差，如 B_0 的标准误差为 0.0792。C 的标准误差为 0.2815 等。

Approximate 95 % Confidence Interval：近似的 95%置信区间。

左栏(Lower)：95%置信区间的下限。

右栏(Upper)：95%置信区间的上限。

注意：95%置信区间的上下限不经过坐标原点则可拒绝系数为 0 的原假设。

从图 14.5 逼近的相关矩阵看来，参数 B_0 与 C 的相关系数约为 0.2702264，不大。B_0 与 D 的相关系数约为 0.1538422，也不大。B_0 与 r 的相关系数约为 −0.3037786 也不大。只是参数 C 与 D、C 与 r、D 与 r 的相关系数的绝对值太大(0.7 左右)。

如果参数之间的相关系数具有很大的“正相关”或很小的“负相关”(如|0.6|左右)，则该模型很可能“超参数”(Over Parameterized)。超参数暗示着一个模型有若干参数不拟合数据。

但出现超参数时不一定意味着模型完全不拟合数据，因为本例的数据量太小，可能导致不足以估计全部参数。

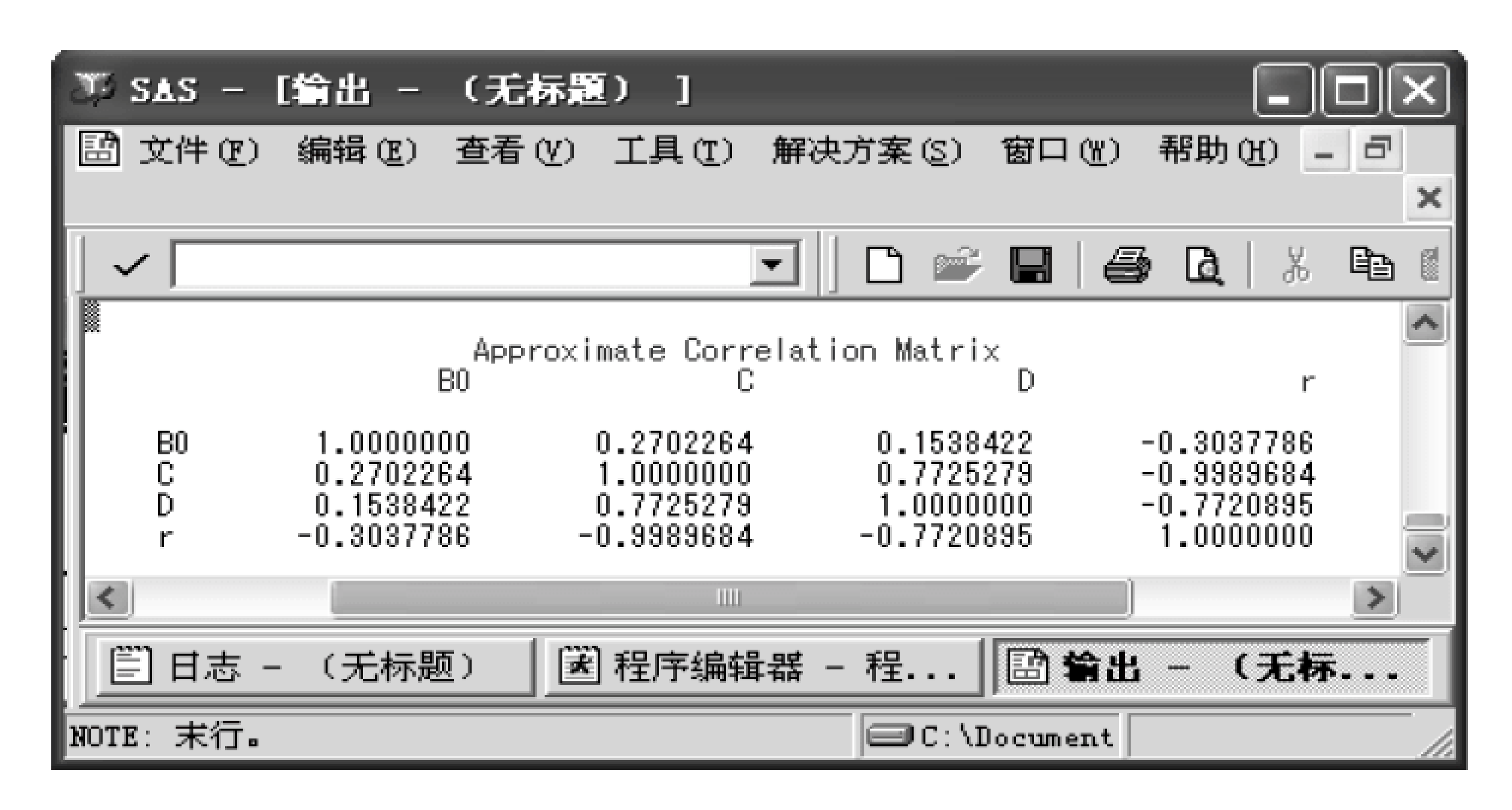


图 14.5 逼近的相关矩阵

最后，根据图 14.4 可“挖掘出”产量的预测公式如下：

$$LOGQ = B_0 + C \times \mathrm{Ln}(D \times L^r + (1-D) \times K^r)$$

$$= 0.0939 - 0.3739 \times \mathrm{Ln}(0.658 \times L^{-2.6999} + 0.342 \times K^{-2.6999})$$

其中，残差平方和为 1.6492(很小，合格)。

说明：同样是这些数据，但用 SPSS 10.1 进行非线性回归时，获得产量的预测公式如下。

$$LOGQ = B_0 + C \times \mathrm{Ln}(D \times L^r + (1-D) \times K^r)$$

$$= 0.1229 + (-0.3398) \times \mathrm{Ln}(0.6617 \times L^{-2.98} + 0.3383 \times K^{-2.98})$$

由此表明，各种统计软件之间也有一些误差，但结果类同。

14.3 拟合抛物线的多项式回归

在社会调查、市场分析和医学研究中，多项式回归分析(拟合抛物线)应用得也很广泛。

14.3.1 多项式回归分析的原始数据

下面图 14.6 中的数据是某最高学府 5 届年龄范围在 18～22 岁的男女生的平均体重，要求建立男生生长发育的曲线(见 14.3.2 节)。

Report
学生体重

AGE	Mean	N	Std. Deviation
18	104.43	7	10.42
19	111.16	37	17.82
20	115.24	144	17.70
21	119.41	109	20.58
22	124.32	38	17.18
Total	116.95	335	18.88

图 14.6 18～22 岁男女生各个年龄组的平均体重

14.3.2　多项式回归的方程式

根据图 14.6 的数据，拟建如下的多项式回归方程式：

$$Y = B_0 + B_1 X + B_2 X^2 \quad (14.2)$$

14.3.3　多项式回归的 SAS 程序

根据图 14.6 中的数据及其公式(14.2)，建立的 SAS 命令文件见程序 14.2。

程序 14.2：拟合 18～22 岁大学生各个年龄组平均体重的回归程序。

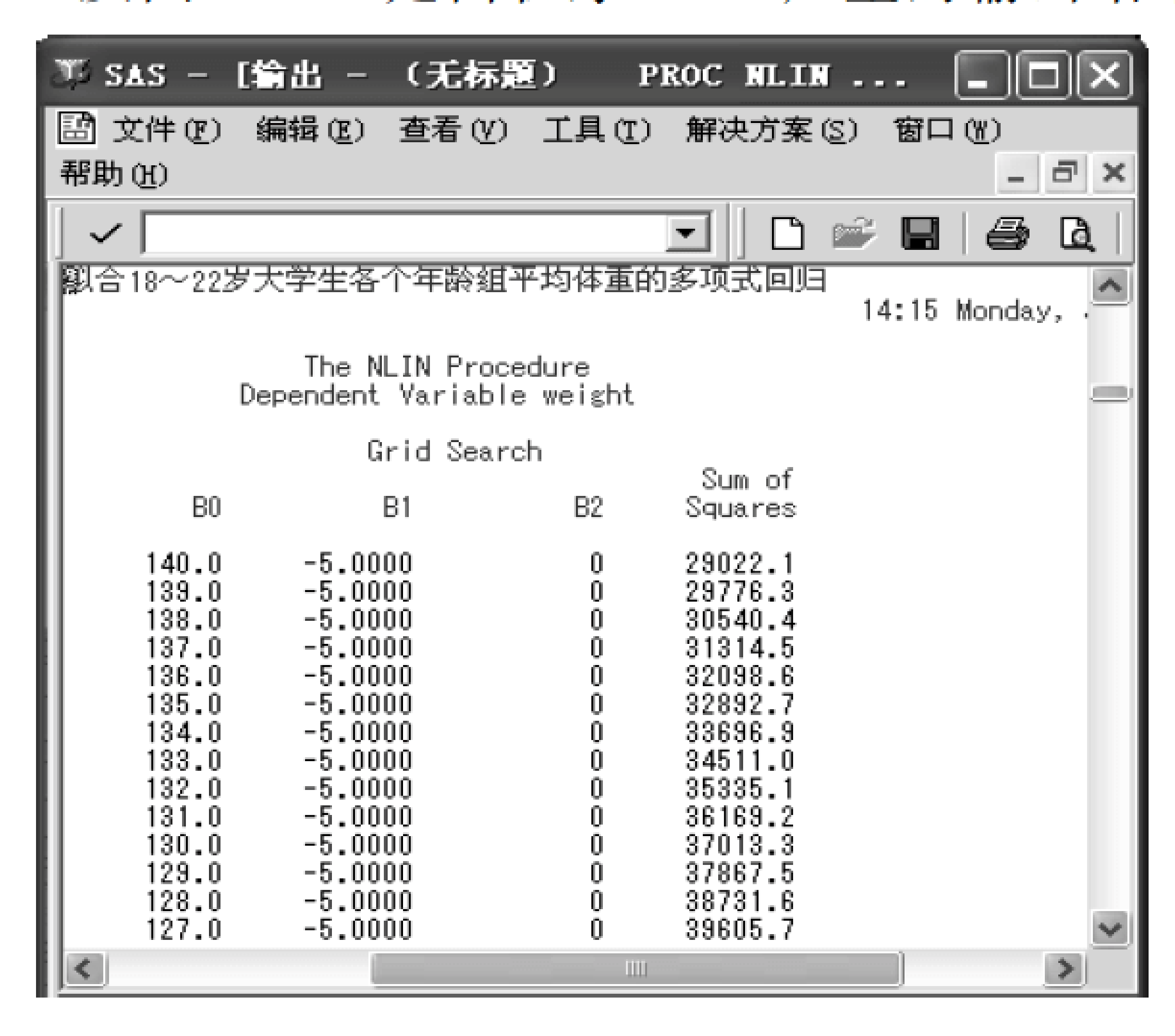
```
DATA WEIGHT;
TITLE                '拟合 18~22 岁大学生各个年龄组平均体重的多项式回归';
INPUT age weight @@ ;
CARDS;
18 104.43 19 111.16 20 115.24 21 119.41 22 124.32
;
PROC NLIN BEST=100 METHOD=MARQUARDT;
PARMS B0 110 TO 140
      B1 - 15 TO - 5
      B2 0 TO 4;
MODEL weight=B0+B1*age+B2*age*age;
```

运行程序 14.2 生成图 14.7～图 14.10 及图 14.12 所示的结果。

14.4　多项式回归的结果与分析

14.4.1　多项式回归的输出结果

图 14.7 至图 14.10 及图 14.12 是由程序 14.2 产生的输出结果。



拟合18～22岁大学生各个年龄组平均体重的多项式回归
14:15 Monday,

The NLIN Procedure
Dependent Variable weight

Grid Search

B0	B1	B2	Sum of Squares
140.0	-5.0000	0	29022.1
139.0	-5.0000	0	29776.3
138.0	-5.0000	0	30540.4
137.0	-5.0000	0	31314.5
136.0	-5.0000	0	32098.6
135.0	-5.0000	0	32892.7
134.0	-5.0000	0	33696.9
133.0	-5.0000	0	34511.0
132.0	-5.0000	0	35335.1
131.0	-5.0000	0	36169.2
130.0	-5.0000	0	37013.3
129.0	-5.0000	0	37867.5
128.0	-5.0000	0	38731.6
127.0	-5.0000	0	39605.7

图 14.7　检测回归系数的初始值

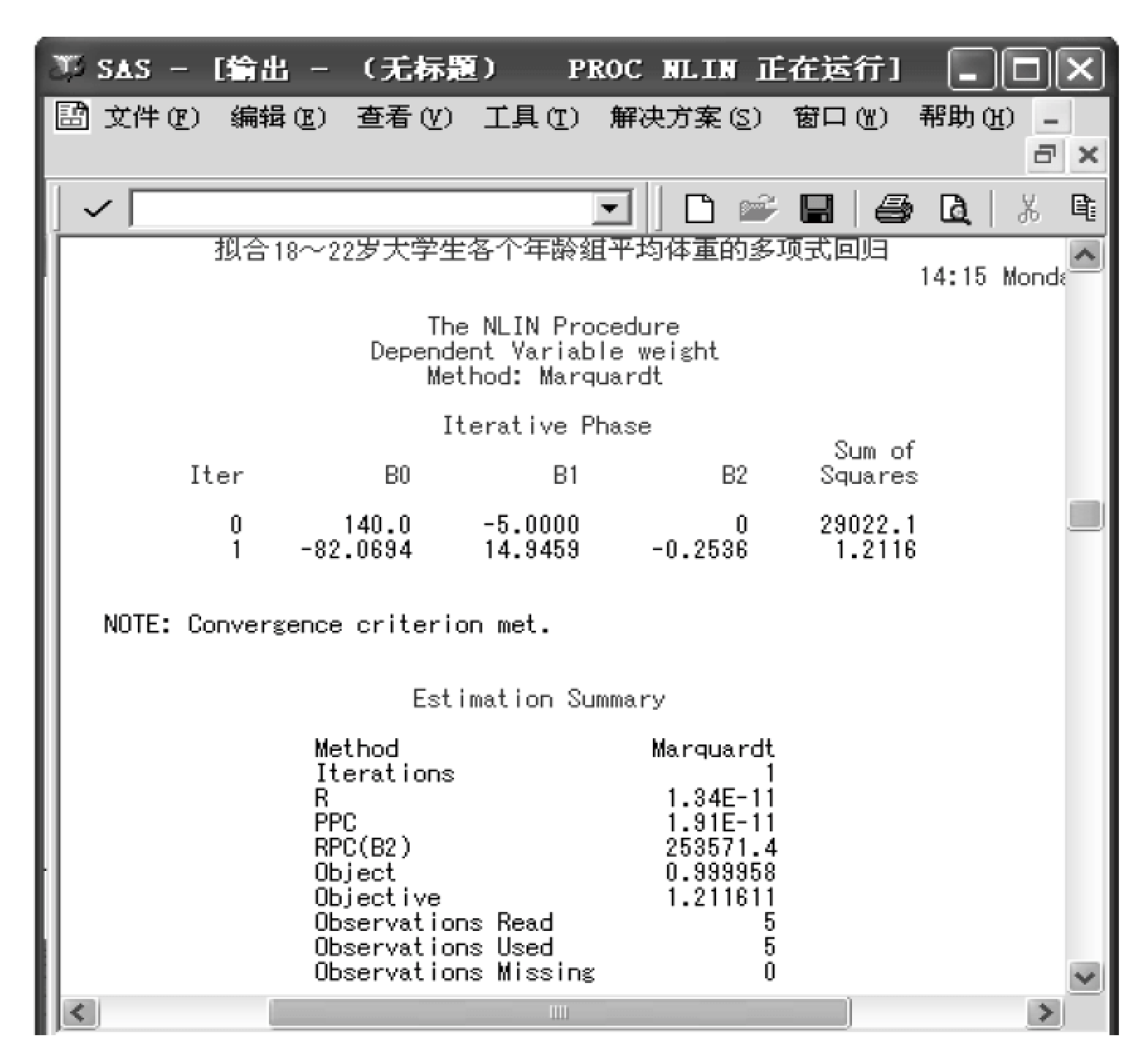


图 14.8　迭代次数为 1 次

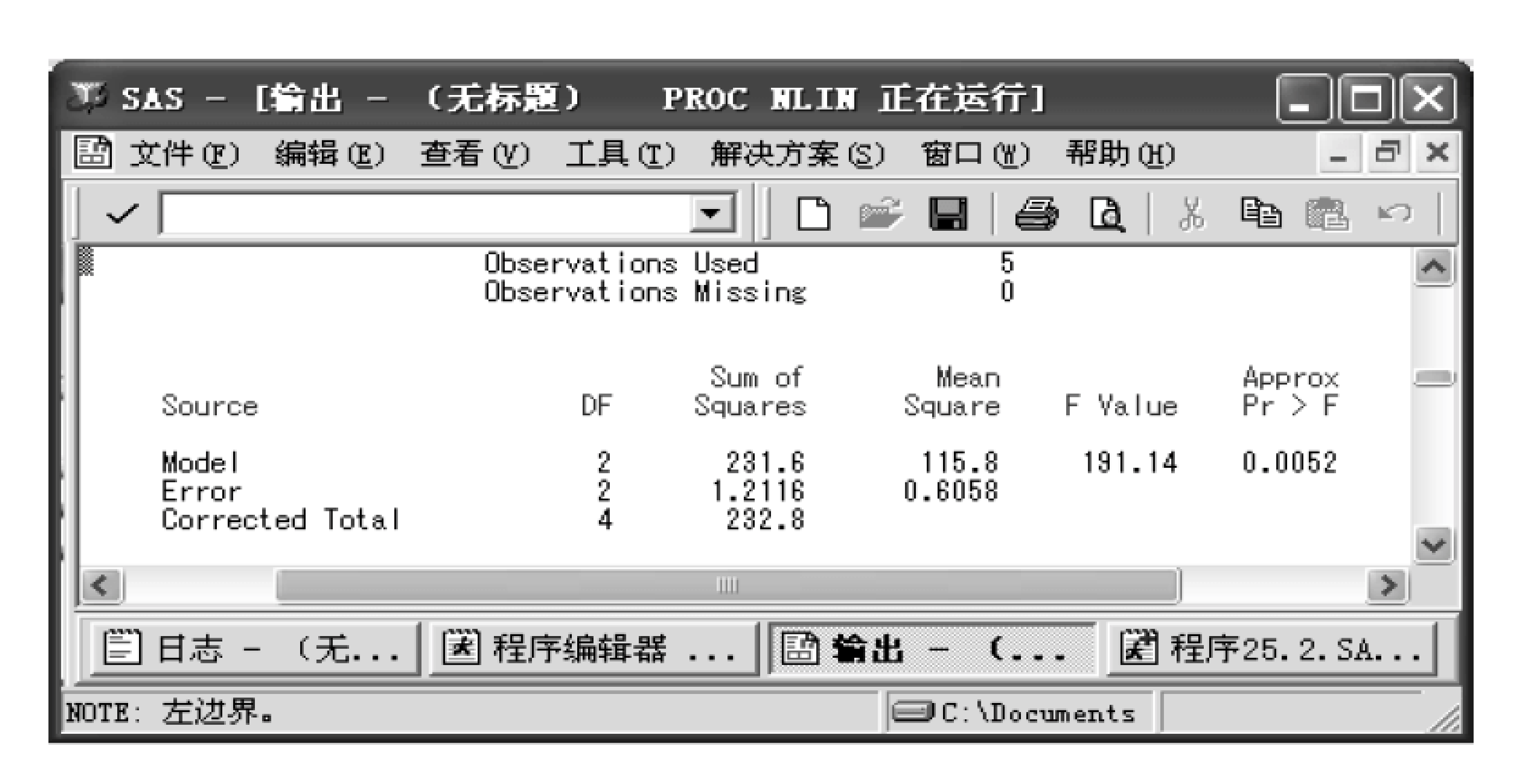


图 14.9　被解释的方差

从图 14.9 看：由回归解释的平方和是 231.6。残差平方和(未被解释的平方和)为 1.2116。所以，回归判定系数为：

$$R^2 = 1-(\text{Residual SS})/(\text{Corrected Total SS})$$
$$=1-1.2116\div 232.8\approx 0.99(\text{接近于 } 1)$$

说明二次项回归曲线拟合度极好。

从图 14.10 看：B0～B2 的 95%置信区间均通过坐标原点，所以 B0～B2 系数均不合格，应另选模型。但是为了继续向下讲授回归模型，假设 B0～B2 系数是有意义的。

由图 14.10 和图 14.11 看出，多项式回归系数 B0、B1、B2 值，无论用 SAS 或是用



图 14.10 由 SAS 计算出的 B0、B1、B2 参数的估计值

SPSS 系统中的曲线回归,所得的值相同。

Independent: age

Dependent	Mth	Rsq	d. f.	F	Sigf	B0	B1	B2
WEIGHT	LIN	.991	3	327.71	.000	18.8520	4.8030	
WEIGHT	QUA	.995	2	191.14	.005	—82.069	14.9459	—.2536
WEIGHT	CUB	.995	2	191.14	.005	—82.069	14.9459	—.2536

Notes:
9 Tolerance limits reached; some dependent variables were not entered.

图 14.11 由 SPSS 计算出的 B0、B1、B2 参数的估计值

表明程序 14.2 中的数据可以选择 METHOD=MARQUARDT 法。



图 14.12 逼近的 B0、B1、B2 相关矩阵

从图 14.12 逼近的相关矩阵看来,参数 B0 与 B1,B0 与 B2,B1 与 B2 三对系数的相关系数约为|0.99|以上,非常大。

说明:参数之间的相关系数的绝对值 0.99 太大,该模型很可能是“超参数”(Over Parameterized),原因是本例的观察值只有 9 个以致不足以估计全部参数。但是回归分析的方法具有普遍意义。

14.4.2 改用“分析家”对话框法进行多项式回归

下面拟用“分析家”对话框法,替代 14.3.3 节编程法并进行多项式回归。

(1) 运行程序 14.2 中的命令与数据,产生 SAS 数据集 Work.weight。

(2) 选择图 14.13 中 SAS 主菜单的"解决方案"→"分析"命令,鼠标指针移到图 14.13带有阴影标记的"分析家"命令上。

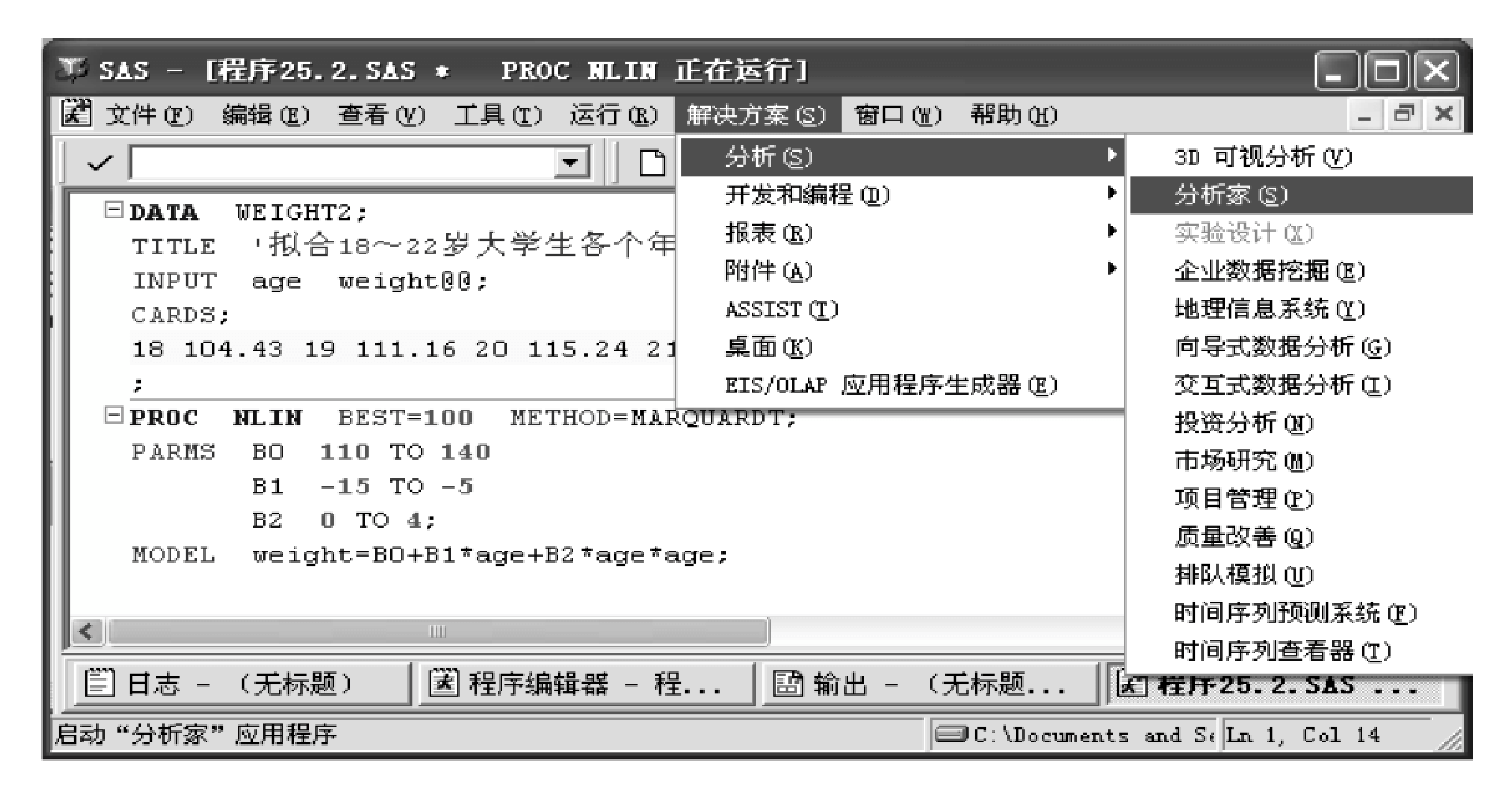


图 14.13 分析家的菜单位置

(3) 选择"分析家"→"文件"→"按 SAS 名称打开"命令,进入图 14.14 后用"下箭头"查找"逻辑库"中的 Work.weight 数据集。



图 14.14 Work.weight 工作文件

(4) 选择文件名 Weight 后单击"确定"(OK)按钮,显示图 14.15 所示的 Work.weight 数据集数据。

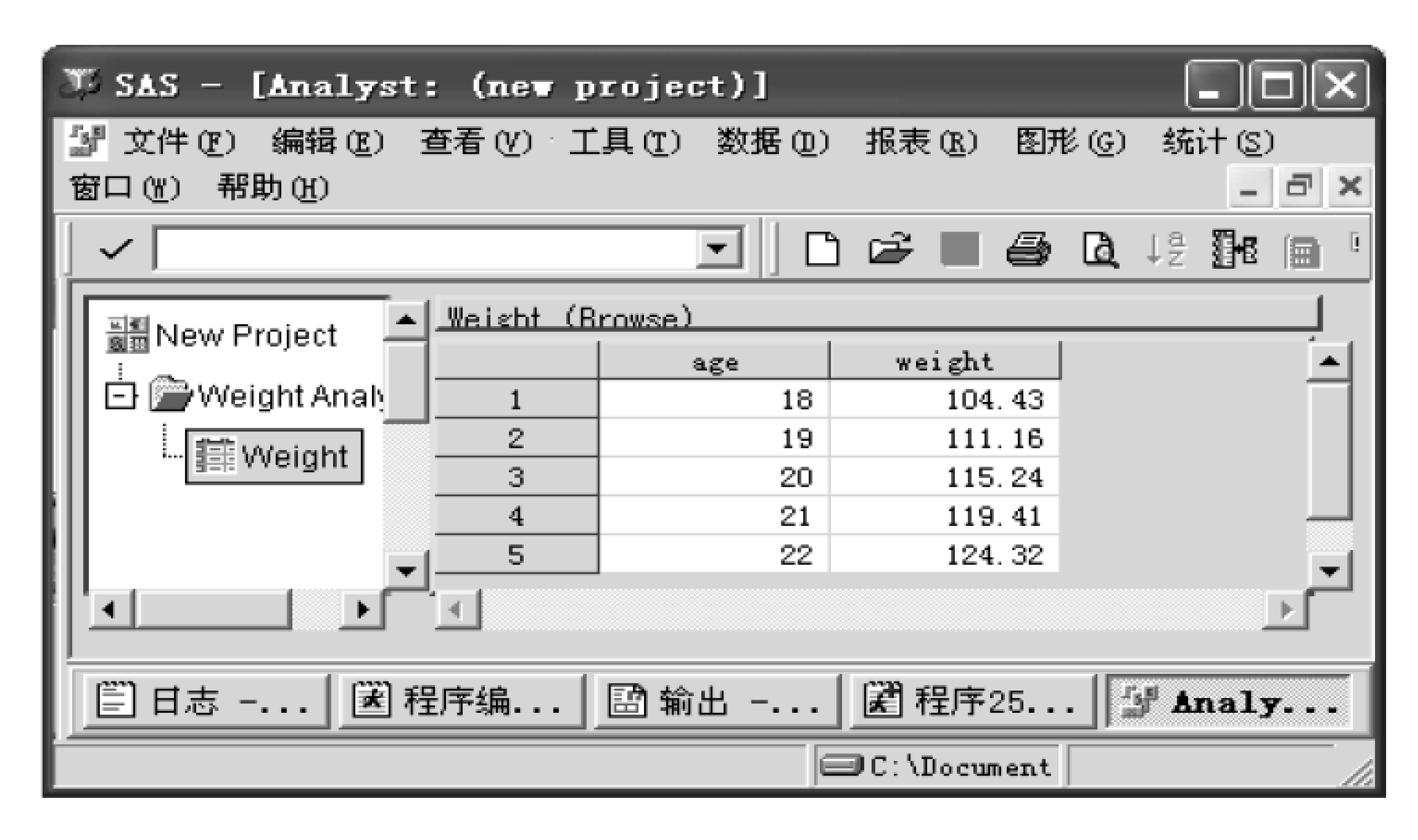


图 14.15 Work.weight 文件的内容(部分)

(5) 选择"统计"→"回归"命令，SAS 显示图 14.16。

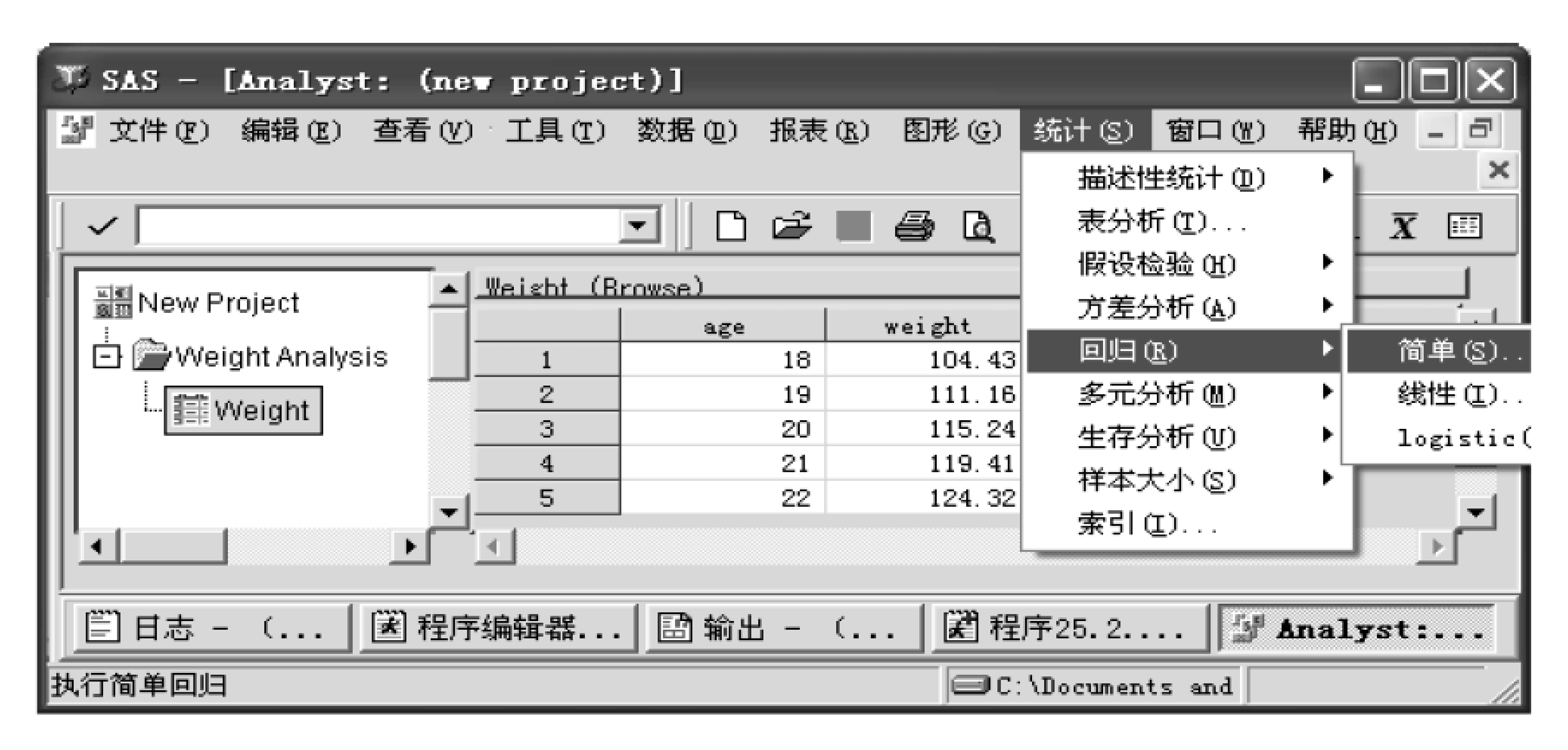


图 14.16 Regression 的菜单位置

(6) 选择图 14.16 中的"简单[S]"命令，进入图 14.17 后选择 Quadratic(二次项回归)及因变量 weight 和自变量 age。

(7) 单击 OK 按钮获得回归输出(见图 14.18)。

回归分析如下。

从图 14.18 的回归分析可以看到：回归系数的检验概率"Pr> |t|"一栏的概率值都大于 α 值 0.05，回归系数不显著，表明二次项模型数据拟合得很不好。如果 Pr> |t| 值小于 α 值 0.05，则二次项模型合格。

所以将图 14.17 中的回归模型改回为 Linear(线性模型)，输出结果见图 14.19 所示。

从图 14.19 的回归分析可以看到：age 回归系数的 Pr> |t| 一栏的概率值 0.0004 小于 α

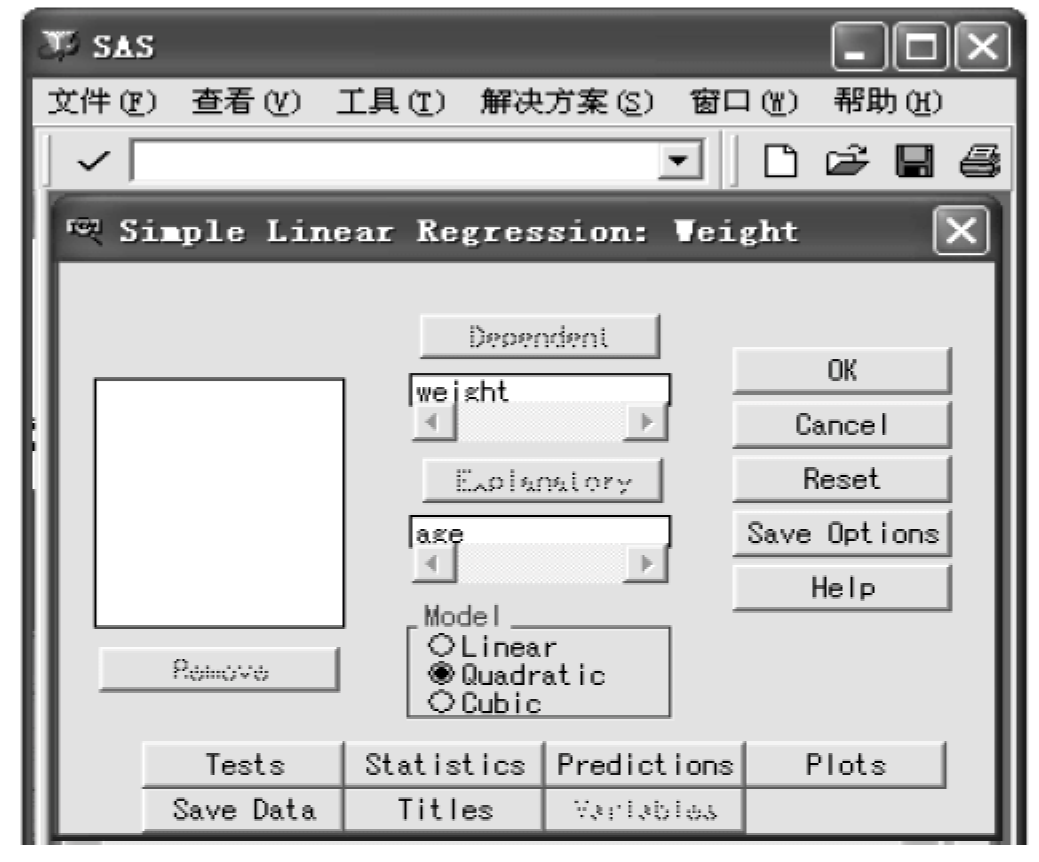
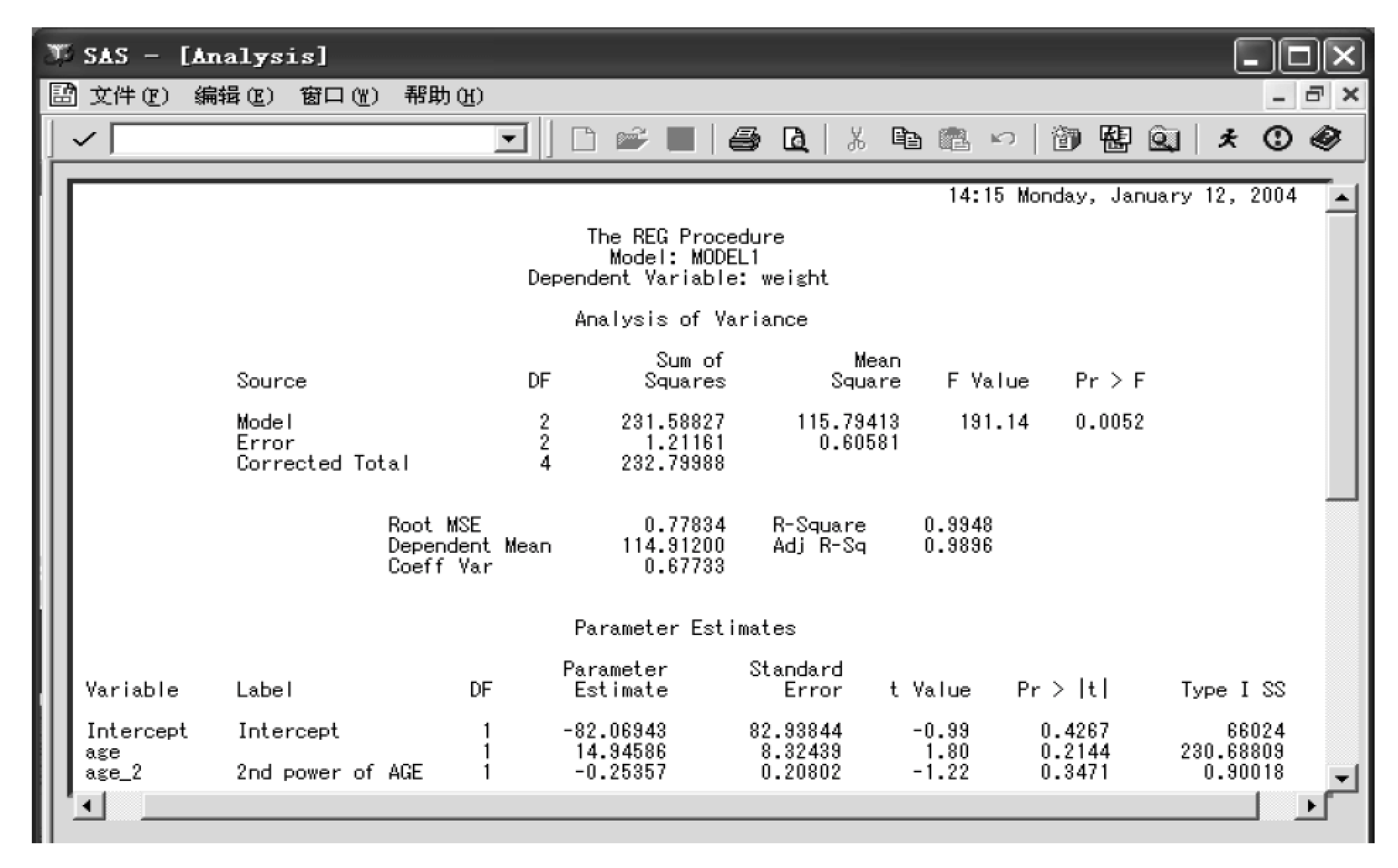


图 14.17 选择 Quadratic(二次项回归)及其因变量 weight 和自变量 age



SAS - [Analysis]

14:15 Monday, January 12, 2004

The REG Procedure
Model: MODEL1
Dependent Variable: weight

Analysis of Variance

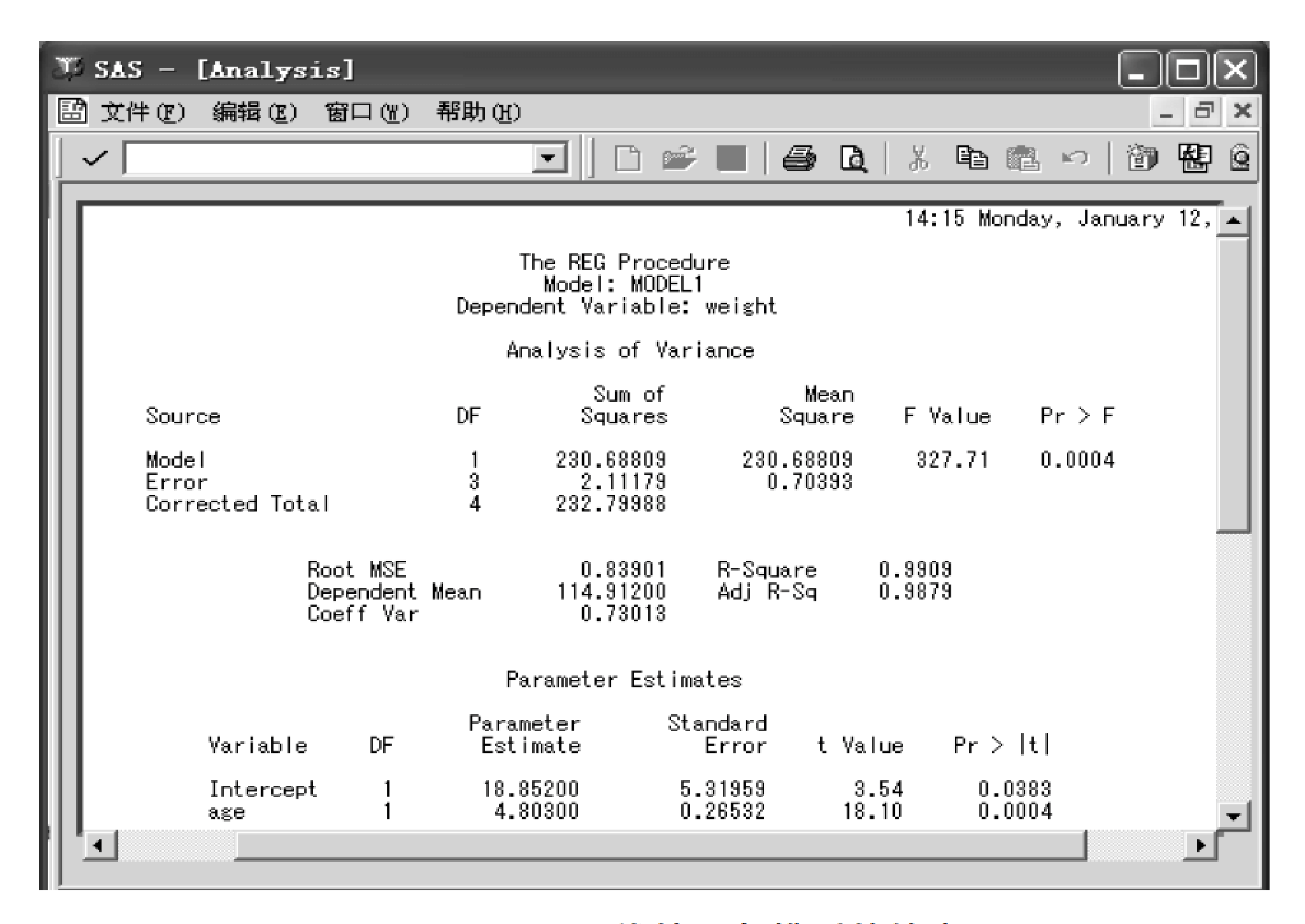
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	231.58827	115.79413	191.14	0.0052
Error	2	1.21161	0.60581		
Corrected Total	4	232.79988			

Root MSE	0.77834	R-Square	0.9948
Dependent Mean	114.91200	Adj R-Sq	0.9896
Coeff Var	0.67733		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS
Intercept	Intercept	1	-82.06943	82.93844	-0.99	0.4267	66024
age		1	14.94586	8.32439	1.80	0.2144	230.68809
age_2	2nd power of AGE	1	-0.25357	0.20802	-1.22	0.3471	0.90018

图 14.18　二元线性回归模型的输出



SAS - [Analysis]

14:15 Monday, January 12,

The REG Procedure
Model: MODEL1
Dependent Variable: weight

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	230.68809	230.68809	327.71	0.0004
Error	3	2.11179	0.70393		
Corrected Total	4	232.79988			

Root MSE	0.83901	R-Square	0.9909
Dependent Mean	114.91200	Adj R-Sq	0.9879
Coeff Var	0.73013		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	18.85200	5.31959	3.54	0.0383
age	1	4.80300	0.26532	18.10	0.0004

图 14.19　一元线性回归模型的输出

值 0.05，回归系数变成非常显著的了，表明一次项模型非常好地拟合数据。

14.4.3　挖掘大学生生长发育的二次曲线模型

现在假设二次项模型拟合得好，则 18～22 岁大学生生长发育的二次曲线模型如下：

$$\text{weight} = B_0 + B_1 \times \text{age} + B_2 \times \text{age} \times \text{age}$$

$$=-82.06943+14.94586\times age-0.25357\times age^2 \quad (14.3)$$

习 题 14

下面 1、2 题是对数曲线回归习题。

1. 对数曲线回归的数学表达式是什么？多项式回归（拟合抛物线）的数学表达式是什么？

2. 下面的图 14.20 是 18～22 岁男生各个年龄组的平均体重，试建立生长发育曲线。

Report

学生体重

AGE	Mean	N	Std. Deviation
19	122.80	15	20.61
20	127.82	68	15.33
21	129.52	63	18.05
22	129.25	24	16.62
Total	128.21	170	17.00

图 14.20　18～22 岁男生各个年龄组的平均体重

提示：请仿照第 14 章 14.3.3 节中多项式回归的 SAS 程序完成本题作业。

3. 拟合 Logistic 曲线回归习题

下述是一个“产量与劳动力（L）、产量与资金（K）”的非线性回归的经济模型。

（1）非线性回归的数据及程序见第 14 章程序 14.1。

（2）非线性回归模型为：

$$\mathrm{LOGQ}=B_0+C\times\log(D\times(L^r)+(1-D)\times(K^r))$$

参数说明：

B_0：截距　　D：分布参数　　C：斜率，即效率参数　　r：替代参数

（3）请分析图 14.21 的结果。

Nonlinear Regression Summary Statistics　　Dependent Variable LOGQ

Source	DF	Sum of Squares	Mean Square
Regression	4	130.00860	32.50215
Residual	26	1.75613	.06754
Uncorrected Total	30	131.76473	
(Corrected Total)	29	61.28965	

R squared＝1－Residual SS/Corrected SS＝ .97135

图 14.21　非线性回归统计量

（4）请分析图 14.22 的结果。

Parameter	Estimate	Asymptotic Std. Error	Asymptotic 95 % Confidence Interval Lower	Upper
B0	.122915203	.078173035	−.037771771	.283602177
C	−.339783002	.273431578	−.901829660	.222263655
D	.661702568	.134549985	.385131113	.938274022
R	−2.980399350	2.285681968	−7.678685922	1.717887222

图 14.22　非线性回归各参数估计的置信区间

（5）请分析图 14.23 的结果。

Asymptotic Correlation Matrix of the Parameter Estimates

	B0	C	D	R
B0	1.0000	.2979	.1801	−.3282
C	.2979	1.0000	.7816	−.9991
D	.1801	.7816	1.0000	−.7815
R	−.3282	−.9991	−.7815	1.0000

图 14.23　非线性回归分析中的参数估计的相关距阵

（6）请写出产量的预测公式（提示：请参阅 14.2 节）。

第 15 章 chapter 15

非线性回归分析二：Logistic 回归与指数回归

在经济统计学中，产量与资金、产量与劳动力之间的关系；在物理学上，自由落体运动与加速度及时间的关系；在人口学上，人口增长率的计算等；诸如此类的复杂计算，无法直接用线性回归分析法，而必须建立非线性回归模型加以解决。

本章讲述的非线性回归包括拟合 Logistic 曲线回归、负指数生长曲线回归和指数回归三方面的内容。

15.1 Logistic 曲线回归

人口增长的例子是最典型的 Logistic 曲线回归例子。数据引自台湾省 1900 年—1980 年期间每 5 年的总人口数据（见程序 15.1）。并在图 15.1 中编辑程序 15.1。

程序 15.1：

```
TITLE 'NonLinear Regression';
DATA rk;
LABEL year= '年份'  rk= '每 5年人口总数,单位:百万人'  t= '每 5年为一组';
INPUT year 1~ 4 rk 6~ 11 t 13~ 14;
CARDS;
1900 02.846 00
1905 3.123 01
1910 3.300 02
1915 3.570 03
1920 3.758 04
1925 4.147 05
1930 4.679 06
1935 5.316 07
1940 6.077 08
1945 6.557 09
1950 7.556 10
1955 9.078 11
1960 10.792 12
```

```
1965 12.628 13
1970 14.676 14
1975 16.150 15
1980 17.805 16
;
PROC FORMAT;
VALUE tf 0='1900年人口总数'  1='1905年人口总数'
     2='1910年人口总数'    3='1915年人口总数'    4='1920年人口总数'
     5='1925年人口总数'    6='1930年人口总数'    7='1935年人口总数'
     8='1940年人口总数'    9='1945年人口总数'    10='1950年人口总数'
     11='1955年人口总数'   12='1960年人口总数'   13='1965年人口总数'
     14='1970年人口总数'   15='1975年人口总数'   16='1980年人口总数';
FORMAT t=tf.;                    /*将 VALUE 步定义的数值标签,反馈给变量 t*/
PROC Nlin DATA=rk BEST=200 METHOD=gradient;
PARMS a=1.8 TO 2 by 0.05 B=-0.11 TO 0 BY 0.1 C=20;
MODEL Rk=C/(1+EXP(a+b*t));
```

在图 15.1 中编辑程序 15.1。

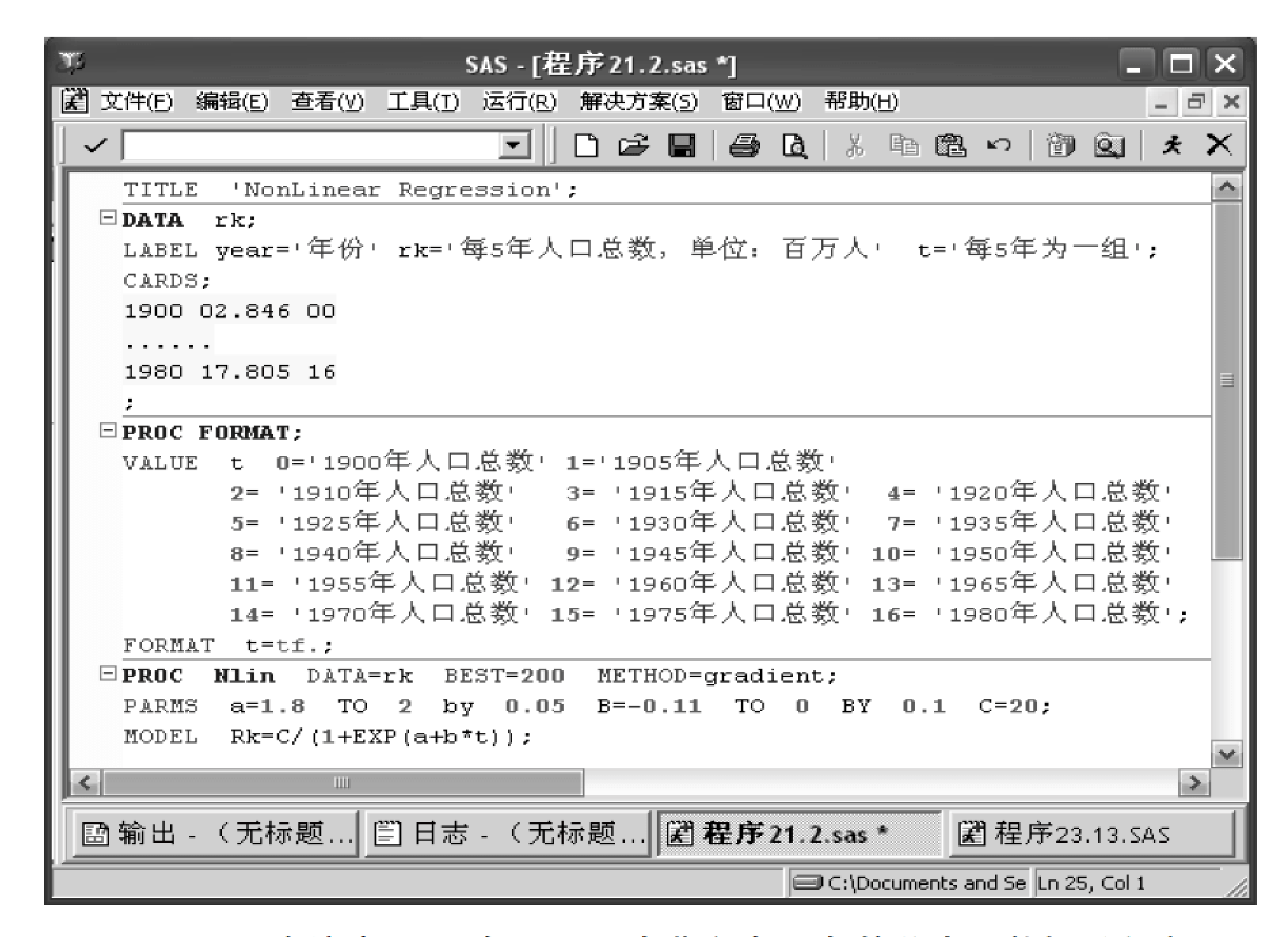


图 15.1　台湾省 1900 年—1980 年期间每 5 年的总人口数据及程序

说明:程序 15.1 中的 METHOD=gradient 语句表明,是采用 Gradient 法,即梯度法(Gradient method),又称急剧下降法(Steepest Descent method)。详见第 14 章的开头介绍。

15.2　从 Logistic 曲线模型解出初始值

1. 建立非线性回归模型

上述图 15.1 的人口总数是随着年份的递增而非线性地增长的，它符合如下人口增长的模型：

$$Y_i = C/(1+e^{A+BTi})+E \tag{15.1}$$

式(15.1)是一个 Logistic 人口增长模型，Y_i 是在 T_i 时间点的人口规模，A 和 B 这两个系数是需要计算的非线性回归方程解的两个初始参数。

式(15.1)中的 C 是表示人口增长模型的渐近线，这条渐近线是由现行数据中任意选出的，既可选择最后一年的人口数，也可选择其中某一年的人口数，本例选择了数据最后一年的总人口数。

式(15.1)中的 E 是误差项，误差的大小依赖于人口数量的变化，但为了计算上的方便，先假定它是常数 0。

2. 拟合 Logistic 曲线回归中初始值 *A*、*B*、*C* 的解

在指数曲线回归之前，首先必须人工计算出初始值 A 和 B，然后将 A、B 的值代入式(15.1)，计算出各个时间点的 Y 值。

为了计算初始值 A、B，首先应给渐近线 C 赋予一个具体的值，本例选择了最后一年即 1980 年台湾省的总人口数并取最接近的整数为 20 百万人。同时假设初始时间点 Y_i=1900 年，即初始的人口总数为 2.846 百万人，因此式(15.1)变换为

$$2.846 = 20 \div (1+e^{A+BTi}) \tag{15.2}$$

然后根据式(15.2)计算 A 和 B 参数值。

1) 先求出系数值 A

A 的值可在 T_0 时间点直接求出，即

$$2.846 = 20 \div (1+e^{A+B*0})$$

移项并整理后得：

$$e^A = 6.027 \tag{15.3}$$

对式(15.3)两边求自然对数后得：

$$A = \text{Ln}(6.027) = 1.8$$

2) 再解出 B 系数值

同理，可从原始数据中 T=1 时间点的人口总数 2.846 百万人中，解出 B 值。

将 A=1.8 代入式(15.2)得：

$$3.123 = 20 \div (1+e^{1.8+B})$$

移项并整理后得：

$$e^{1.8+B} = 5.4 \tag{15.4}$$

式(15.4)两边取对数得：

$$1.8 + B = \text{Ln}(5.4) \tag{15.5}$$

$$B = 1.69 - 1.8 = -0.11$$

3. 把 *A*、*B*、*C* 系数代入 Logistic 模型

由于目前的 SAS 版本不能用对话框进行非线性回归分析，因此需要在程序 15.1 的最后增加指数回归命令，详见程序 15.2。

1）指数回归的数据及命令行

程序 15.2：

```
TITLE 'NonLinear Regression';
DATA rk2;
LABEL year='年份'  rk='每 5年人口总数,单位:百万人'  t='每 5年为一组';
INPUT year 1~4 rk 6~11 t 13~14;
CARDS;
1900 02.846 00
1905 3.123 01
1910 3.300 02
1915 3.570 03
1920 3.758 04
1925 4.147 05
1930 4.679 06
1935 5.316 07
1940 6.077 08
1945 6.557 09
1950 7.556 10
1955 9.078 11
1960 10.792 12
1965 12.628 13
1970 14.676 14
1975 16.150 15
1980 17.805 16;
PROC FORMAT;
VALUE tF 0='1900年人口总数'  1='1905年人口总数'
      2='1910年人口总数'    3='1915年人口总数'    4='1920年人口总数'
      5='1925年人口总数'    6='1930年人口总数'    7='1935年人口总数'
      8='1940年人口总数'    9='1945年人口总数'    10='1950年人口总数'
      11='1955年人口总数'   12='1960年人口总数'   13='1965年人口总数'
      14='1970年人口总数'   15='1975年人口总数'   16='1980年人口总数';
FORMAT t=tF.;
PROC Nlin DATA=rk METHOD= GRADIENT;
PARMS A=1.8 to 2 by 0.1 B=-0.11 TO 0.01 by 0.1 C=20;
MODEL Rk=C/(1+EXP(a+b*t));
RUN;
```

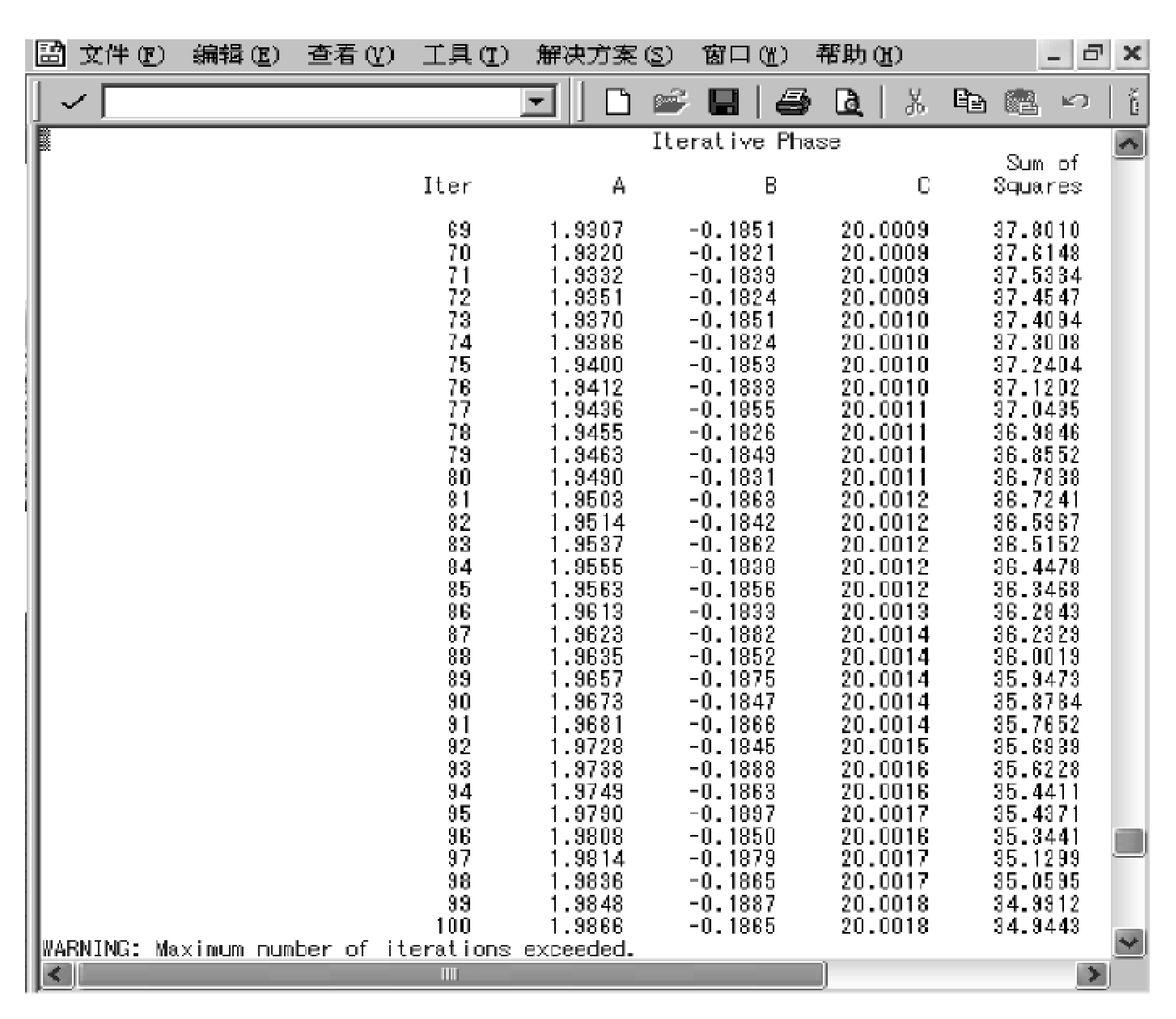
2）生成 SAS 的数据集 Work.rk2

（1）在程序编辑器窗口选择“运行”→“提交”命令，运行程序 15.2，形成 SAS 数据集 Work.Rk2。

（2）与此同时产生图 15.2～图 15.5 所示的指数回归的输出结果。

15.3 拟合 Logistic 曲线回归的分析

运行程序 15.2 产生图 15.2～图 15.5 所示的输出结果。

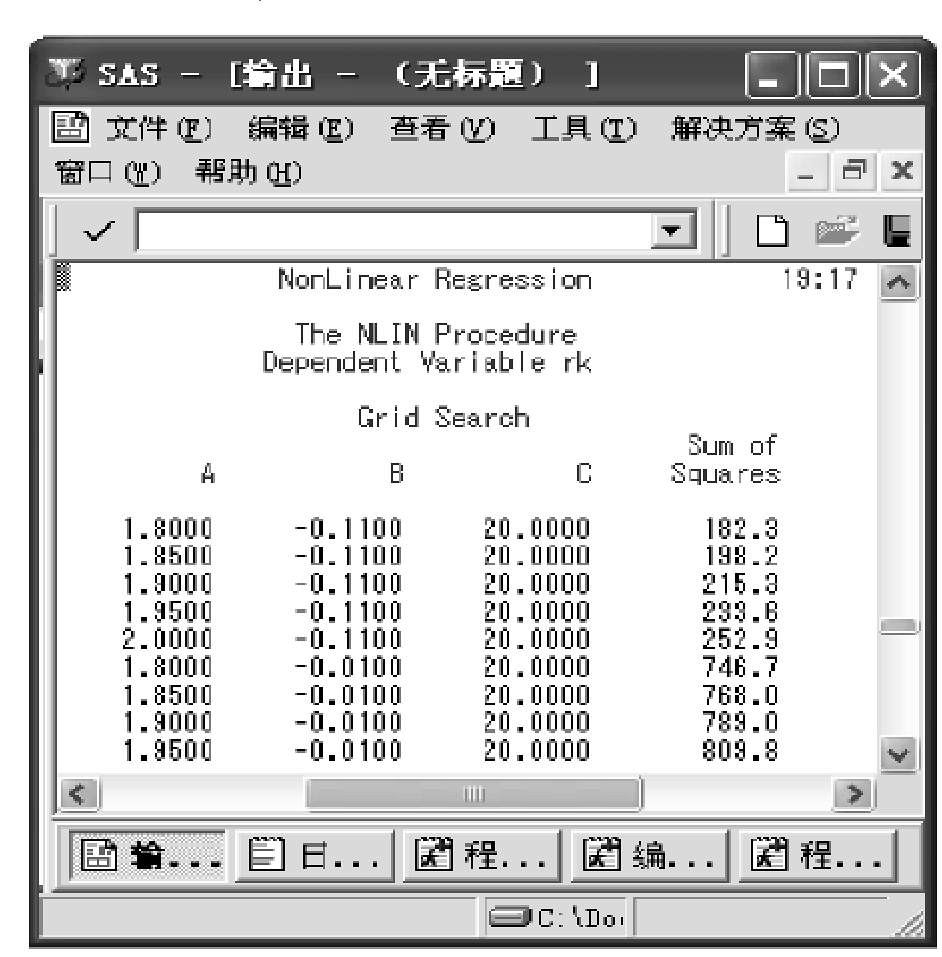


Iterative Phase

Iter	A	B	C	Sum of Squares
69	1.9307	-0.1851	20.0009	37.8010
70	1.9320	-0.1821	20.0009	37.6148
71	1.9332	-0.1839	20.0009	37.5334
72	1.9351	-0.1824	20.0009	37.4547
73	1.9370	-0.1851	20.0010	37.4094
74	1.9386	-0.1824	20.0010	37.3008
75	1.9400	-0.1853	20.0010	37.2404
76	1.9412	-0.1838	20.0010	37.1202
77	1.9436	-0.1855	20.0011	37.0435
78	1.9455	-0.1826	20.0011	36.9846
79	1.9463	-0.1849	20.0011	36.8552
80	1.9490	-0.1831	20.0011	36.7838
81	1.9503	-0.1868	20.0012	36.7241
82	1.9514	-0.1842	20.0012	36.5967
83	1.9537	-0.1862	20.0012	36.5152
84	1.9555	-0.1838	20.0012	36.4478
85	1.9563	-0.1856	20.0012	36.3468
86	1.9613	-0.1833	20.0013	36.2843
87	1.9623	-0.1882	20.0014	36.2329
88	1.9635	-0.1852	20.0014	36.0019
89	1.9657	-0.1875	20.0014	35.9473
90	1.9673	-0.1847	20.0014	35.8784
91	1.9681	-0.1866	20.0014	35.7652
92	1.9728	-0.1845	20.0015	35.6939
93	1.9738	-0.1888	20.0016	35.6228
94	1.9749	-0.1863	20.0016	35.4411
95	1.9790	-0.1897	20.0017	35.4371
96	1.9808	-0.1850	20.0016	35.3441
97	1.9814	-0.1879	20.0017	35.1299
98	1.9836	-0.1865	20.0017	35.0595
99	1.9848	-0.1887	20.0018	34.9912
100	1.9866	-0.1865	20.0018	34.9443

WARNING: Maximum number of iterations exceeded.

(a) 每步的残差平方和



NonLinear Regression　19:17

The NLIN Procedure
Dependent Variable rk

Grid Search

A	B	C	Sum of Squares
1.8000	-0.1100	20.0000	182.3
1.8500	-0.1100	20.0000	198.2
1.9000	-0.1100	20.0000	215.3
1.9500	-0.1100	20.0000	233.6
2.0000	-0.1100	20.0000	252.9
1.8000	-0.0100	20.0000	746.7
1.8500	-0.0100	20.0000	768.0
1.9000	-0.0100	20.0000	789.0
1.9500	-0.0100	20.0000	809.8

(b) 回归参数 A、B、C

图 15.2　每步的残差平方和及残差估计

15.3.1 参数估计

1. 何时迭代终止

图 15.2(a)中,每步的残差平方和分别为:

0　　182.3:(未显示出来)表示第 0 次(即初始)的残差平方和为 182.3。

98　　35.0595:表示第 90 次迭代时的残差平方和为 35.0595。

99　　34.9912:表示第 99 次迭代时的残差平方和为 34.9912。

100　34.9443:表示第 100 次迭代时的残差平方和为 34.9443。

说明:迭代次数越往后,残差平方和则越小,最后两次迭代时残差平方和如果几乎接近,则达到了收敛标准,迭代终止。

本例迭代了 100 次(默认值为最大迭代次数 100 次),虽然尚未收敛,但是不得不终止迭代。

图 15.2(b)中,指出了回归参数 A、B、C 的具体内容。

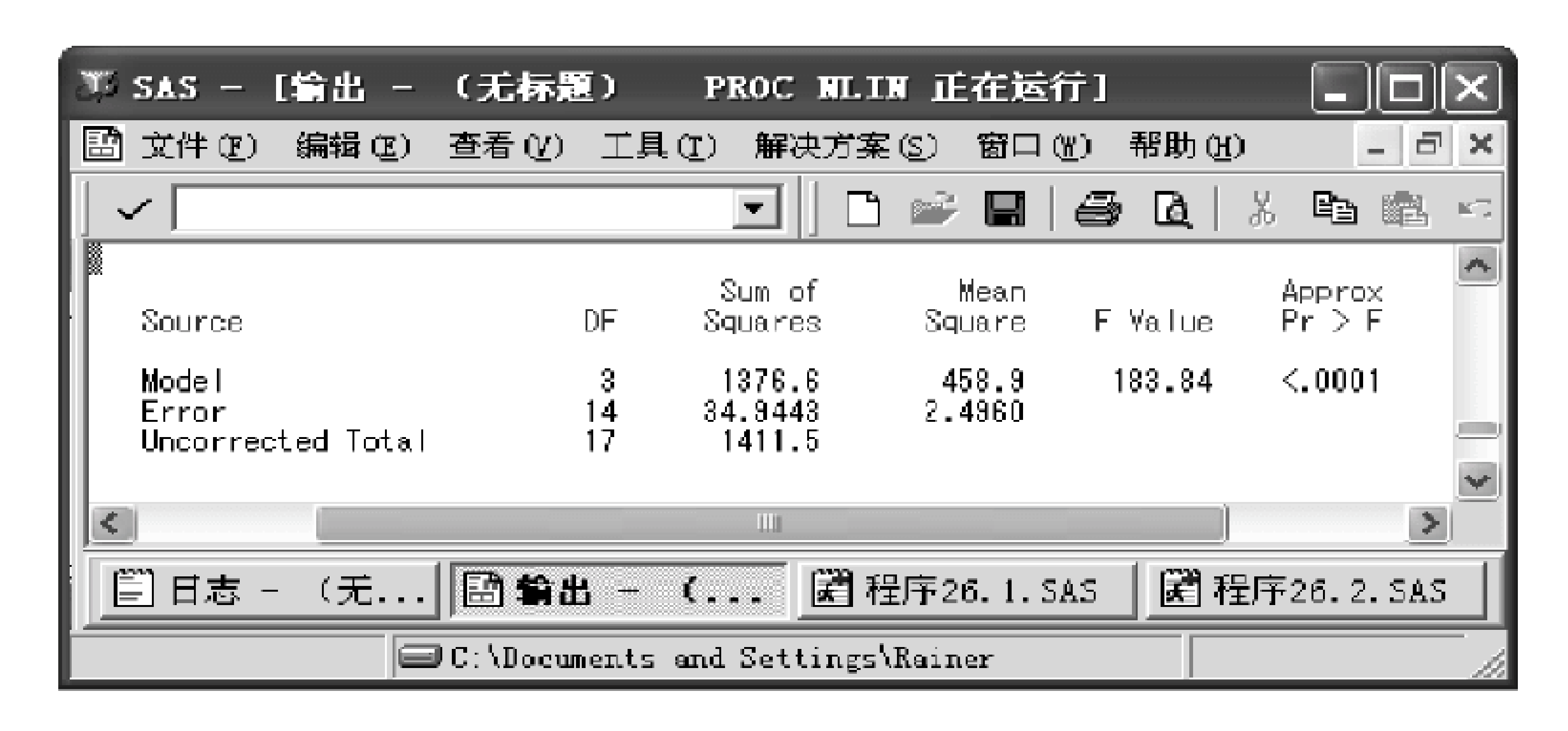
A:相当于回归方程中的截距,数值 1.8 是用户解出的初始值。

B:相当于回归方程中的斜率;数值(−0.11)是用户解出的初始值。

C:回归方程的渐近线;C=20 也是人为解出的初始值。

2. 非线性回归模型的统计量

图 15.3 是非线性回归模型的概括统计量。



Source	DF	Sum of Squares	Mean Square	F Value	Approx Pr > F
Model	3	1376.6	458.9	183.84	<.0001
Error	14	34.9443	2.4960		
Uncorrected Total	17	1411.5			

图 15.3　非线性回归模型的概括统计量

3. 统计量解释

Model Sum of Squares:是已被解释的回归平方和(其值为 1376.6)。

Erorr Sum of Squares:未被解释的残差平方和(其值为 34.9443)。

Uncorrected Total Sum of Squares:因变量的总平方和(其值为 1411.5)。

Corrected Total Sum of Squares:偏离均值的平方和,SAS 9 不显示此值。

R squared=1−Residual SS/Corrected SS=1−0.03=0.97:确定系数,或称判定系

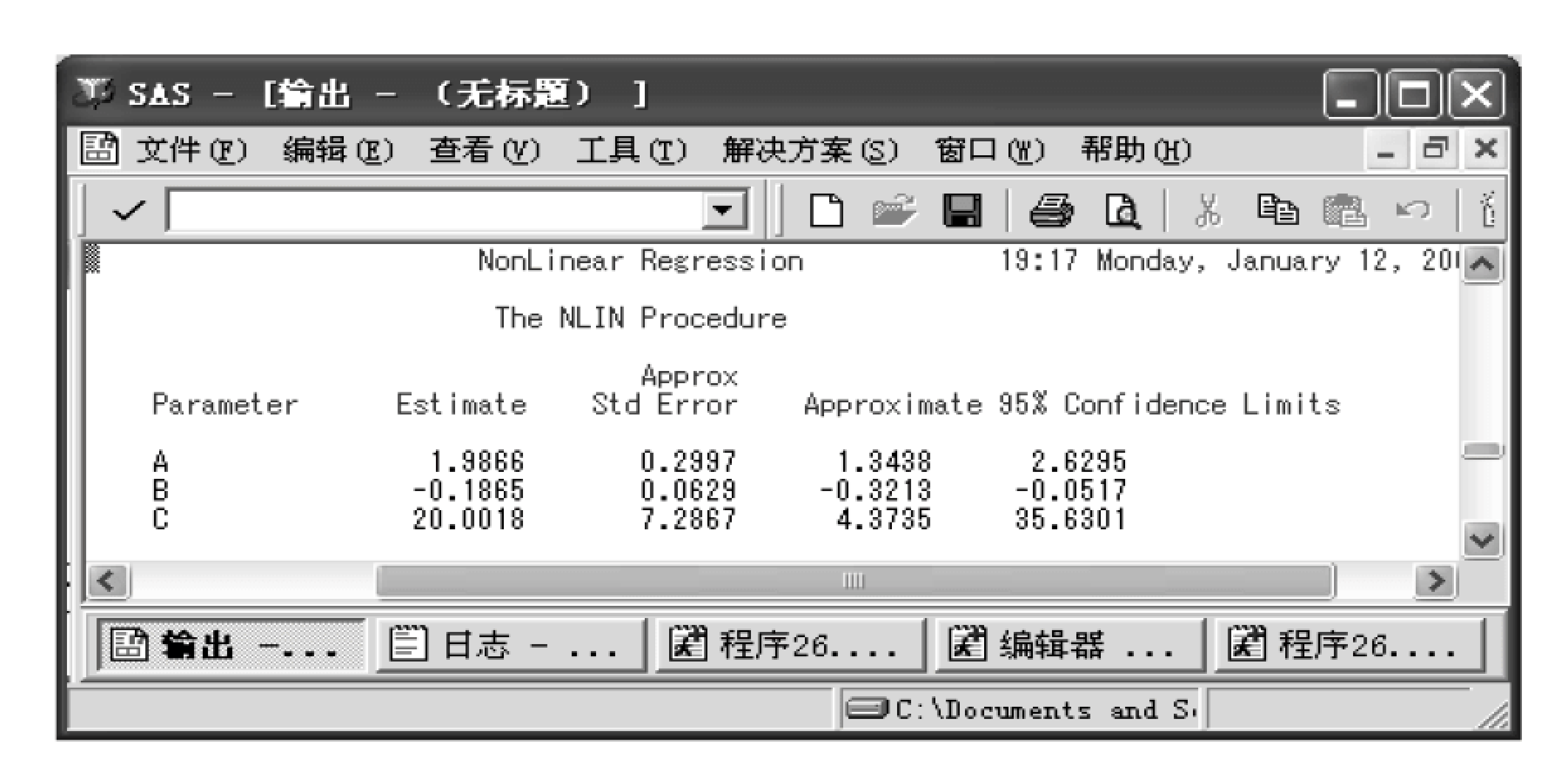
数。0.97 值表示被模型解释的方差占因变量(Rk,人口)的总方差的 97%。

本例中(R squared)=0.97 接近于 1,说明模型很好地拟合了数据。

15.3.2 参数近似的置信区间

1. 置信区间

在非线性回归模型中,不能获得各个参数精确的置信区间,因此,必须如图 15.4 所示对样本采取“渐近线”的近似计算。



Parameter	Estimate	Approx Std Error	Approximate 95% Confidence Limits	
A	1.9866	0.2997	1.3438	2.6295
B	-0.1865	0.0629	-0.3213	-0.0517
C	20.0018	7.2867	4.3735	35.6301

图 15.4 参数近似的置信区间

对图 15.4 的各个参数说明如下:

Parameter:参数,有 A、B、C 三个参数。

Estimate:参数的估计值;如 A 参数的估计值为 1.9866,B 参数的估计值为 −0.1865 等。

Approx Std Error:逼近的标准误差。如渐近线 C 的标准误差为 7.2867,截距 A 的标准误差为 0.2997 等。

Approximate 95% Confidence Limits:逼近的 95%置信度。

Lower:95%置信度的下限,如 A 参数的下限为 1.3438。

Upper:95%置信度的上限,如 A 参数的上限为 2.6295。

从 95%置信度的上下限值看来,A、B、C 三个参数的上下限均不经过坐标原点 0,所以有理由拒绝原假设。接着继续观察下面的“2”。

2. 逼近的相关矩阵(见图 15.5)

说明:图 15.5 中的 B 与 C 参数之间的相关系数具有很大的相关系数,该模型很可能“超参数”(Over Parameterized)。超参数暗示着模型的 B 与 C 参数不拟合数据。原因可能在于数据量太小致使不足以估计全部参数。

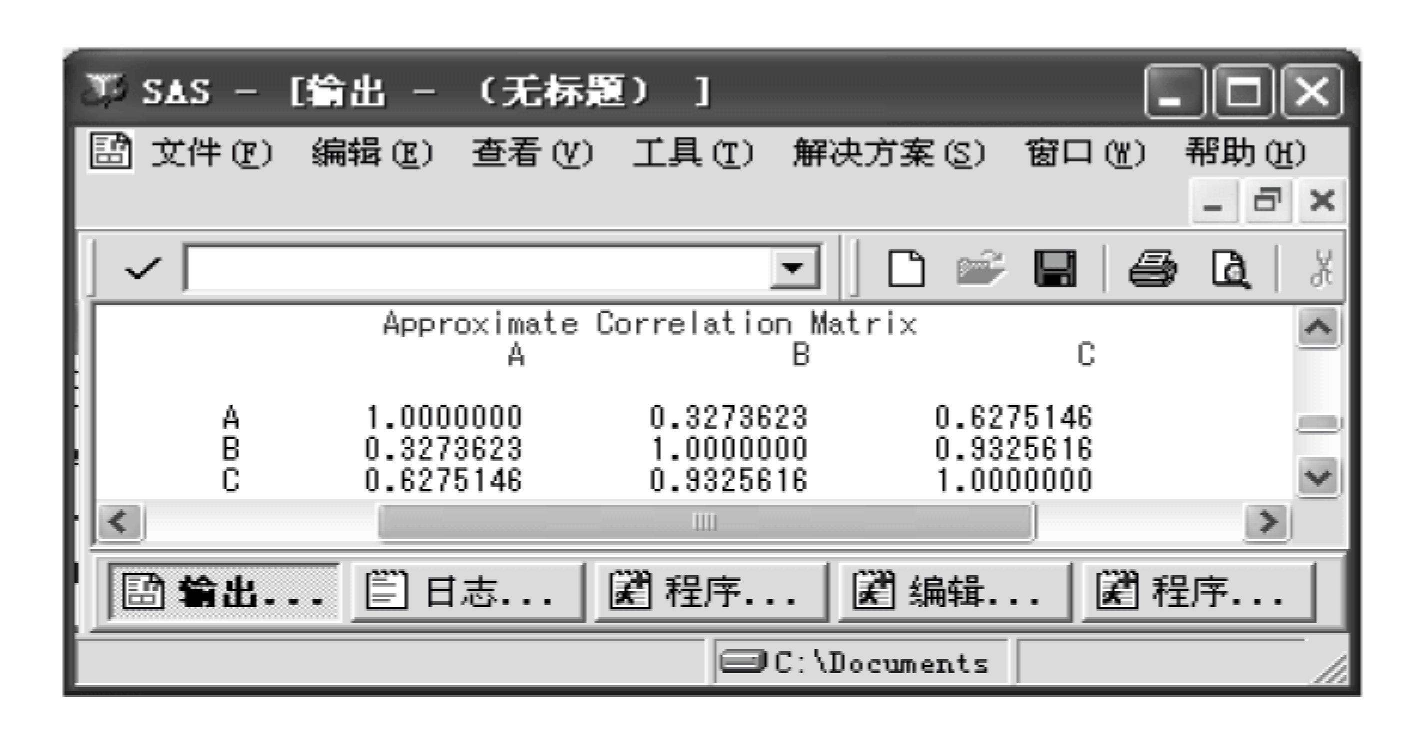


图 15.5 逼近的相关矩阵

3. 残差验证

回归时应存储期望值和残差，并画出残差对应年份的曲线图(如散点图等)，便于研究模型的拟合优度。

计算期望值和逼近的标准误差一般是用界外点(Outlier)检测法，以及对影响点(个案)的分析法，即用线性回归过程。残差则是从非线性回归模型中计算的。从图 15.3 中的 Residua 值 34.9443 看，残差平方和为略为偏大，但合格。

15.3.3 用 Logistic 曲线发掘人口数据

由图 15.4 统计与预测结果可以得出台湾省 2005 年人口的预测值如下。

$$Y = C/(1+\mathrm{EXP}(A+B\times T)) = 20/(1+\mathrm{EXP}(1.9866-0.1865\times T))$$
$$=20/(1+\mathrm{EXP}(1.9866-0.1865\times 21))$$
$$\approx 20$$ 百万人(但迭代到 100 次后尚未收敛)

说明：

(1) 加上误差项，可预测到台湾省 2005 年人口的预测值约为 21(百万人)。

(2) 同样的数据，用程序 15.2 中的 METHOD=GRADIENT 回归法，比用其他回归法相对要准确。

(3) 同样的数据，分别用 SPSS 和 SAS 中的相同回归方法时，预测的结果会有很大的误差。根据试验，之所以误差如此大，原因之一是因为迭代到 100 次后尚未收敛。

15.4 负指数生长曲线回归

1. 什么时候应采用负指数生长曲线模型

符合下面式(15.6)的数据应采用负指数生长曲线模型。

$$Y = B_0(1-\mathrm{e}^{-B1\times X}) \tag{15.6}$$

2. 负指数生长曲线的例子

负指数生长的数据例子见表 15.1，假定数据拟合负指数生长。

表 15.1　某公司每月产品的合格率

X(月份)	1	2	3	4	5	6
Y(合格率)	0.69	0.73	0.76	0.80	0.84	0.89
X(月份)	7	8	9	10	11	12
Y(合格率)	0.95	1.00	0.95	0.89	1.00	0.98

本节着重按表 15.2 所示的体重减肥数据，挖掘体重减肥趋势的预测模型。

表 15.2　体重减肥趋势（累计数）

X(月份)	1	2	3	4	5	6
Y(斤)	6.9	7.3	7.6	8.0	8.4	8.9
X(月份)	7	8	9	10	11	12
Y(斤)	9.5	9.9	10.5	11.5	12.5	13.5

假定表 15.2 的数据拟合负指数生长趋势，则按照下面所介绍的方法解答。

3. 负指数生长曲线初始值的解法

进行指数曲线回归之前，首先必须计算初始值 B_0、B_1 和 X，然后将初始值代入式(15.6)，计算出各个时间点的 Y 值。

为了计算初始值，首先应给 B_0 赋予一个具体的值，本例选择了最后一个月（即第 12 个月）最接近的体重为 14 斤，同时假设时间点 $Y_i = X_1$ 点的体重为 6.9 斤，因此式(15.6)可变换为：

$$Y = B_0(1 - e^{-B_1 \times X})$$

$$6.9 = 14(1 - e^{-B_1 \times 1}) \tag{15.7}$$

然后根据式(15.7)计算出如下的 B_1 参数值。

1）先求出 B_1 的系数值

B_1 值可在 X_1 时间点直接求出，即

$$6.9 = 14(1 - e^{-B_1 \times 1})$$

移项并整理后得：

$$e^{-B_1 \times 1} = 0.51 \tag{15.8}$$

对式(15.3)两边求自然对数并移项后得：

$$B_1 = 1.67$$

2）再解出 B_0 的系数值

同理，可从原始数据的 $X=2$ 时间点，解出

$$B_0 = 7.6 \tag{15.9}$$

4. "挖掘"负指数生长的曲线模型

将初始值 B_0、B_1 值代入负指数生长的曲线模型后获得式(15.10)。

$$Y = B_0(1 - e^{-B_1 \times X}) = 7.6(1 - e^{-1.67 \times X}) \tag{15.10}$$

下面用简单的编程法定义数据和建立模型,详见程序 15.3。

5. 用编程法产生负指数生长曲线

1) 用编程法(见程序 15.3)发掘负指数生长曲线

程序 15.3:数据及其指数回归的命令。

```
TITLE '负指数生长曲线模型';
LABEL X='月'   Y='累积体重减轻:斤';
DATA YER;
INPUT X Y @@;
TITLE '      ';

CARDS;
1 6.9 2 7.3 3 7.6 4 8.0 5 8.4 6 8.9 7 9.5 8 9.9
9 10.5 10 11.5 11 12.5 12 13.5
;
PROC NLIN DATA=YER BEST=100 METHOD=MARQUARDT;
PARMS b0=7.6 TO 8 b1=1.67 TO 2;
Model Y=B0* (1-EXP(-b1* X));
RUN;
```

2) 运行程序 15.3 产生结果

(1) 在程序编辑器窗口选择 Run(运行)→Submit(提交)命令,运行程序 15.3→形成 SAS 数据集 work. YER。

(2) 与此同时产生图 15.7 至图 15.10 所示的指数回归输出结果。

15.5 分析负指数生长曲线

下面分析图 15.6 至图 15.10 所示的回归结果。

1. 参数估计

1) 何时迭代终止

(1) 图 15.6 是每步迭代时的残差平方和。

对于图 15.6,应该观察每步的残差平方和。

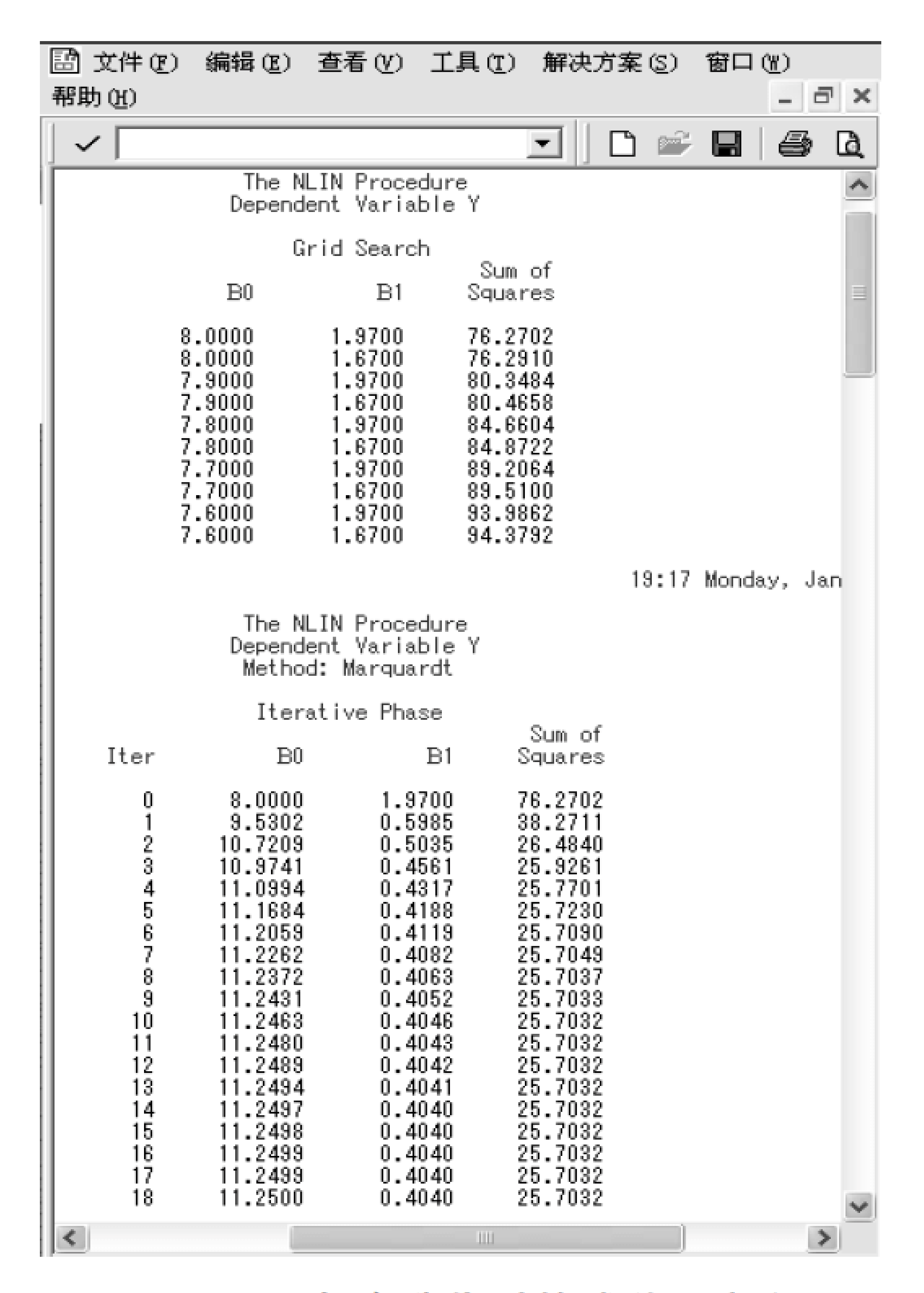
迭代次数越往后,残差平方和则越小,最后两次迭代时残差平方和如果几乎接近(仅差 0.00001 左右),则达到了收敛标准,本例迭代了 18 次就终止。

从图 15.6 还可看出:

B0：相当于回归方程中的截距(8.0000)。

B1：相当于回归方程中的斜率(1.9700)。

(2) 下面图 15.7 是迭代的小结。



The NLIN Procedure
Dependent Variable Y

Grid Search

B0	B1	Sum of Squares
8.0000	1.9700	76.2702
8.0000	1.6700	76.2910
7.9000	1.9700	80.3484
7.9000	1.6700	80.4658
7.8000	1.9700	84.6604
7.8000	1.6700	84.8722
7.7000	1.9700	89.2064
7.7000	1.6700	89.5100
7.6000	1.9700	93.9862
7.6000	1.6700	94.3792

19:17 Monday, Jan

The NLIN Procedure
Dependent Variable Y
Method: Marquardt

Iterative Phase

Iter	B0	B1	Sum of Squares
0	8.0000	1.9700	76.2702
1	9.5302	0.5985	38.2711
2	10.7209	0.5035	26.4840
3	10.9741	0.4561	25.9261
4	11.0994	0.4317	25.7701
5	11.1684	0.4188	25.7230
6	11.2059	0.4119	25.7090
7	11.2262	0.4082	25.7049
8	11.2372	0.4063	25.7037
9	11.2431	0.4052	25.7033
10	11.2463	0.4046	25.7032
11	11.2480	0.4043	25.7032
12	11.2489	0.4042	25.7032
13	11.2494	0.4041	25.7032
14	11.2497	0.4040	25.7032
15	11.2498	0.4040	25.7032
16	11.2499	0.4040	25.7032
17	11.2499	0.4040	25.7032
18	11.2500	0.4040	25.7032



Estimation Summary

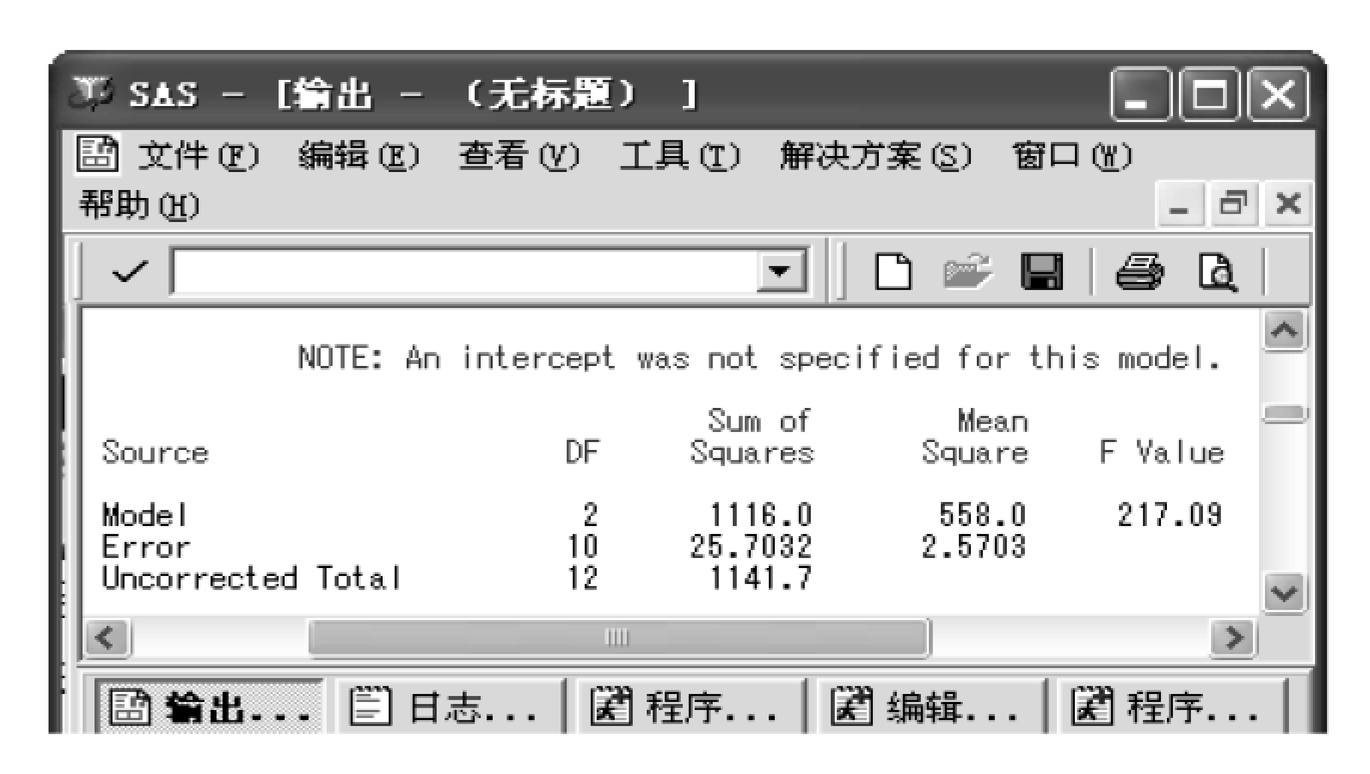
Method	Marquardt
Iterations	18
Subiterations	8
Average Subiterations	0.444444
R	6.248E-6
PPC(B1)	5.472E-6
RPC(B1)	0.00001
Object	2.07E-10
Objective	25.70317
Observations Read	12
Observations Used	12
Observations Missing	0

图 15.6　每步迭代时的残差平方和　　　　图 15.7　迭代的小结

从图 15.7 可看出，本例迭代了 18 次便终止迭代。

2) 非线性回归模型的统计量

如图 15.8 所示，回归平方和为 1116.0。但未被解释的残差平方和为 25.7032，较大，因此引起了：

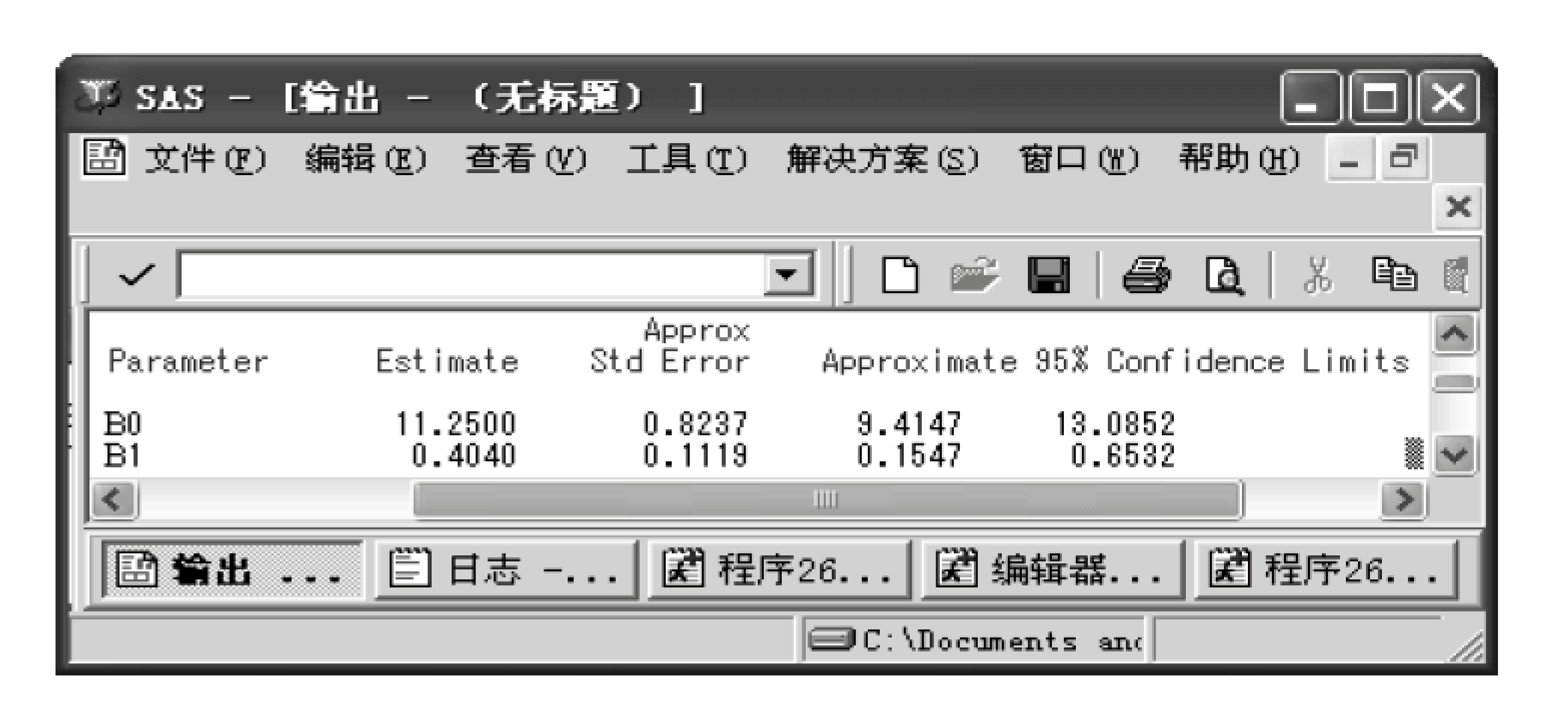


NOTE: An intercept was not specified for this model.

Source	DF	Sum of Squares	Mean Square	F Value
Model	2	1116.0	558.0	217.09
Error	10	25.7032	2.5703	
Uncorrected Total	12	1141.7		

图 15.8　被解释的方差

$R^2=1-$(Residual SS)/(Corrected Total SS)无法计算。



Parameter	Estimate	Approx Std Error	Approximate 95% Confidence Limits	
B0	11.2500	0.8237	9.4147	13.0852
B1	0.4040	0.1119	0.1547	0.6532

图 15.9 回归截距 B_0 与斜率 B_1

图 15.9 各个参数的含义如下。

Parameter：有 B_0、B_1 参数。

Estimate：参数的估计值；如 B_0 参数的估计值为 11.2500。

Approx Std. Error：逼近的标准误差。如 B_0 参数的标准误差为 0.8237(偏大)。

Approximate 95 % Confidence Limits：逼近的 95%置信度。

Lower：95%置信度的下限，如 B_0 下限为 9.4147。

Upper：95%置信度的上限，如 B_0 上限为 13.0852。

从 95%置信度的上下限值看来，两个参数的上下限均不经过坐标原点 0，因此有理由拒绝原假设，即系数不为 0。再看下面的“3”。



Approximate Correlation Matrix

	B0	B1
B0	1.0000000	-0.7552140
B1	-0.7552140	1.0000000

图 15.10 逼近的相关矩阵

3) 逼近的相关矩阵

从图 15.10 逼近的相关矩阵看来，参数 B0 与 B1 的相关系数为|−0.7552140|，比较大。原因之一是样本的数据量太小致使不足以估计全部参数。

4) 残差验证

在 SAS 的回归中，应记住存储期望值和残差，并画出残差对应年份的曲线图(如散点图等)，便于研究模型的拟合优度。

计算期望值和逼近的标准误差一般是用 Outlier 检测法以及对影响点个案的分析法，即用线性回归过程。而残差则是从非线性回归模型中计算的。从图 15.8 知：残差平方和 25.7032。

所以本数据不拟合负指数生长曲线，但是所述的负指数生长曲线回归方法不失一般性。请试用下述 15.6 节的正指数回归重做回归分析，然后比较二者的结果。

2. 负指数生长曲线的模型

根据图 15.9 中的回归截距 B_0 与斜率 B_1，可写出以下的负指数生长曲线的模型：

$$Y = B_0(1 - e^{-B_1 \times X}) = 11.25(1 - e^{-0.404X}) \qquad (15.11)$$

15.6 拟合指数曲线 $Y=Ae^{BX}$ 回归

设有表 15.3 中的数据，试拟合指数曲线 $Y=Ae^{BX}$。

表 15.3 拟合指数曲线 $Y=Ae^{BX}$ 的数据(体重逐月减轻，即非累积)

月　份	1	2	3	4	5	6	7	8
体重减肥(克)	1900	1800	1600	1400	1100	700	500	500

15.6.1 建立指数曲线 $Y=Ae^{BX}$ 的回归模型

根据表 15.3 中的数据，建立的指数曲线 $Y=Ae^{BX}$ 回归模型见程序 15.4。

程序 15.4：

```
DATA jfei;
INPUT month decrease@ @ ;
CARDS;
1 1900 2 1800 3 1600 4 1400 5 1100 6 700 7 500 8 500
;
PROC NLIN BEST=50 METHOD=marquardt;
PARMS A=3000 TO 2600
      B=-0.5 TO 0;
MODEL decrease=a * EXP(b * month);
RUN;
```

运行程序 15.4 产生图 15.11 至图 15.15 所示的结果。

15.6.2 分析指数曲线 $Y=Ae^{BX}$ 回归结果

1. 图 15.11 至图 15.14 是指数曲线 $Y=Ae^{BX}$ 的回归输出图形

2. 结果分析

(1) 方差分析。从图 15.14 可以看到：

Model Sum of Squares (Regression SS)：是已被解释的回归平方和(其值为 13397502)。

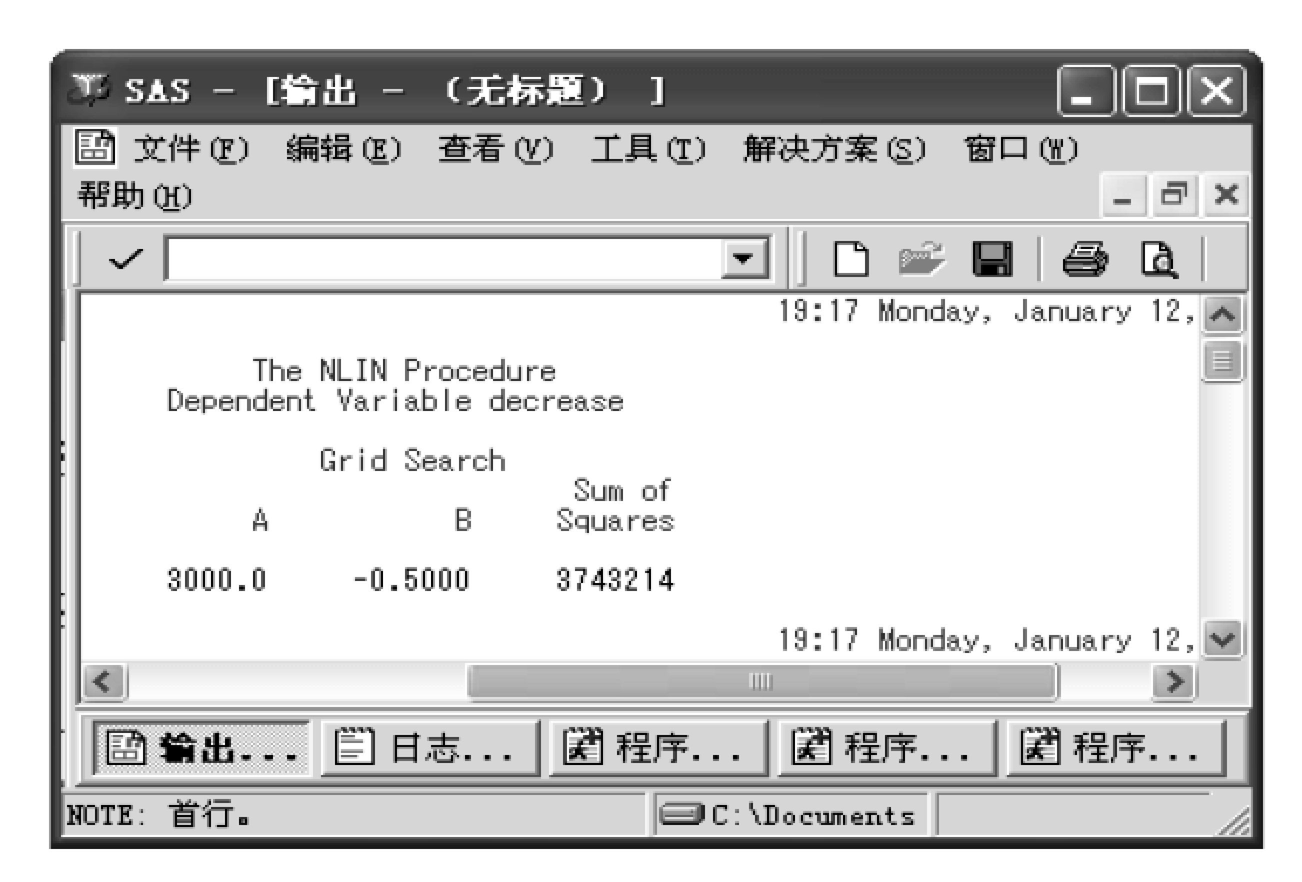


图 15.11 模型中 A、B 系数的初始值

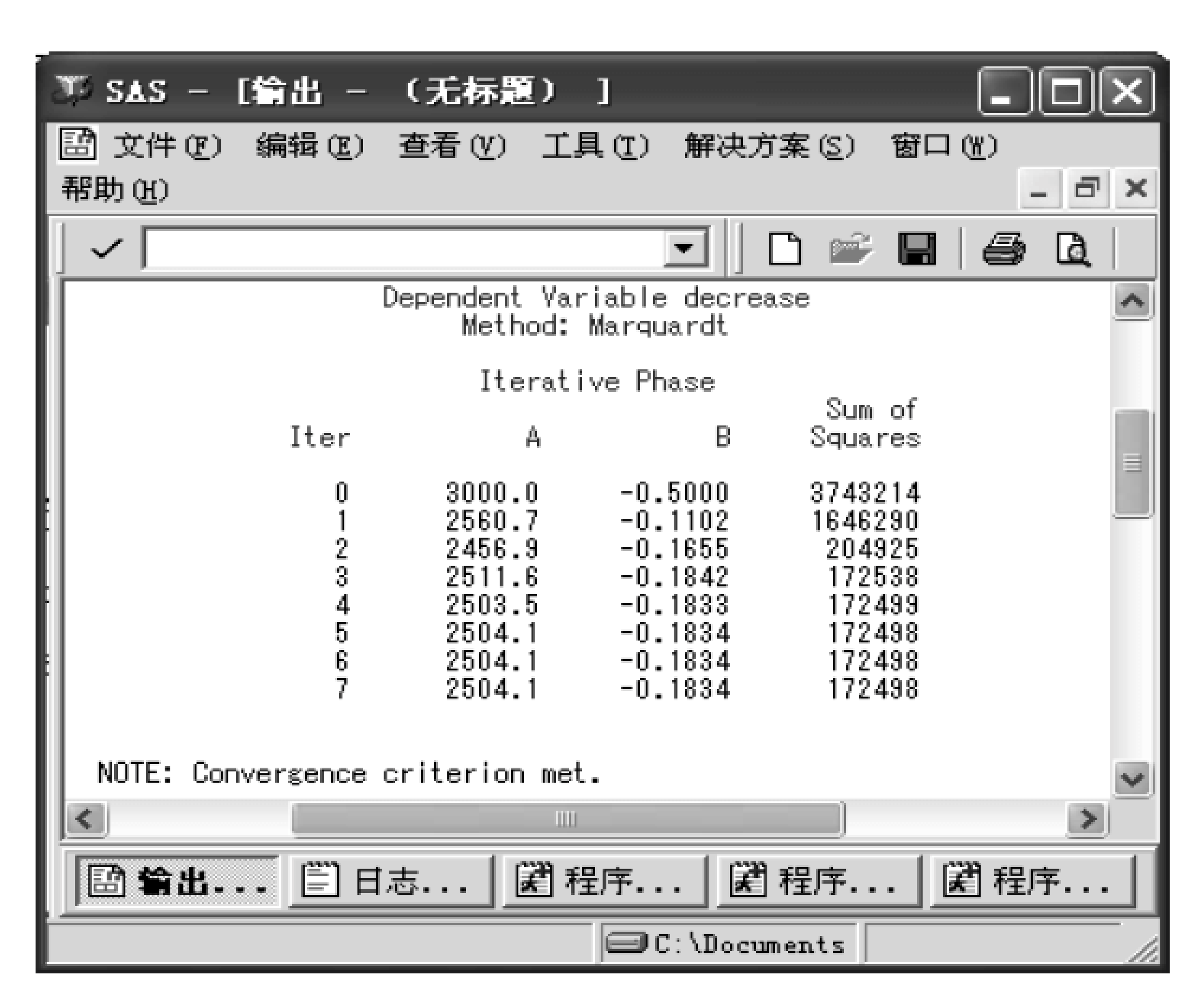


图 15.12 迭代过程

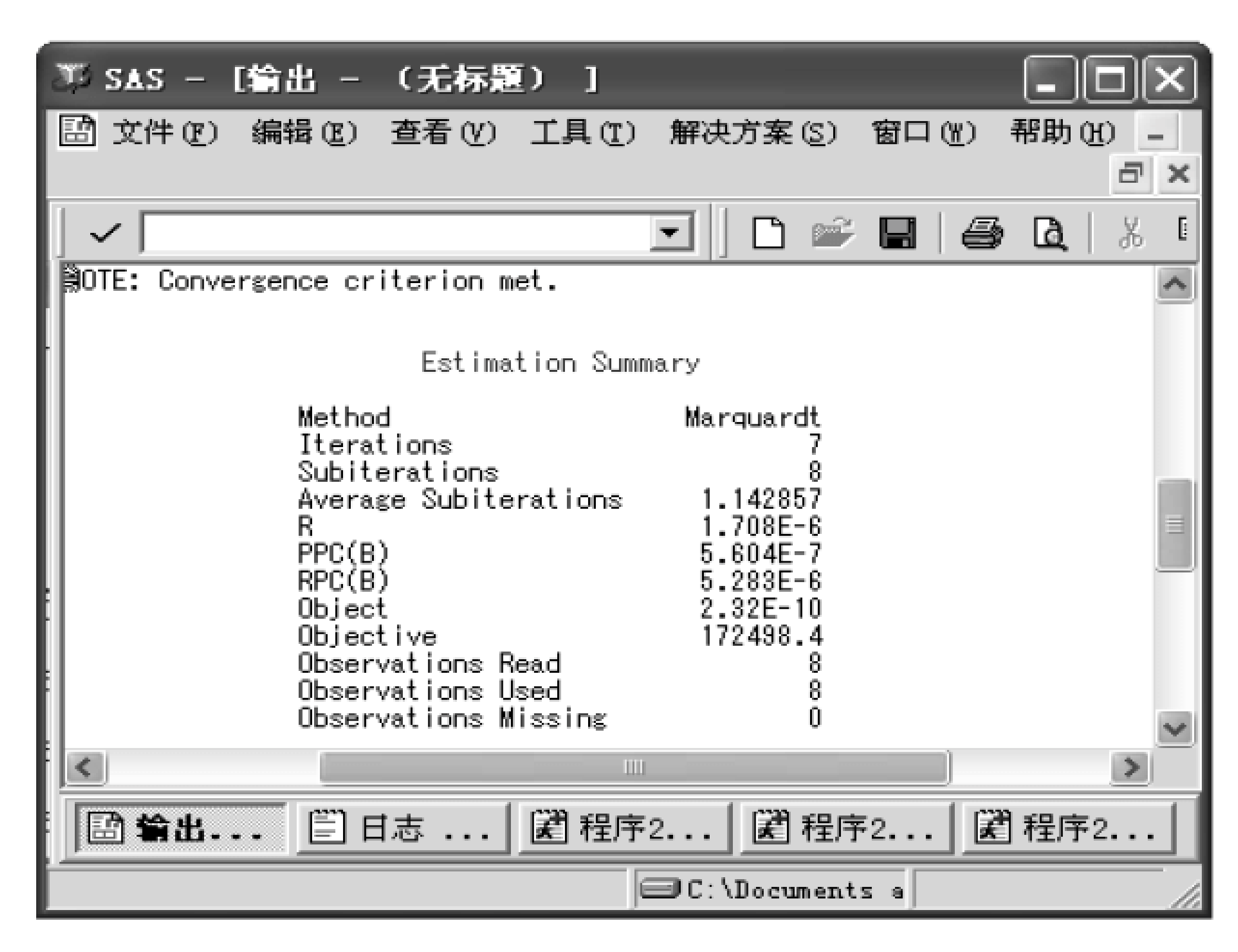


图 15.13 程序运行的概述

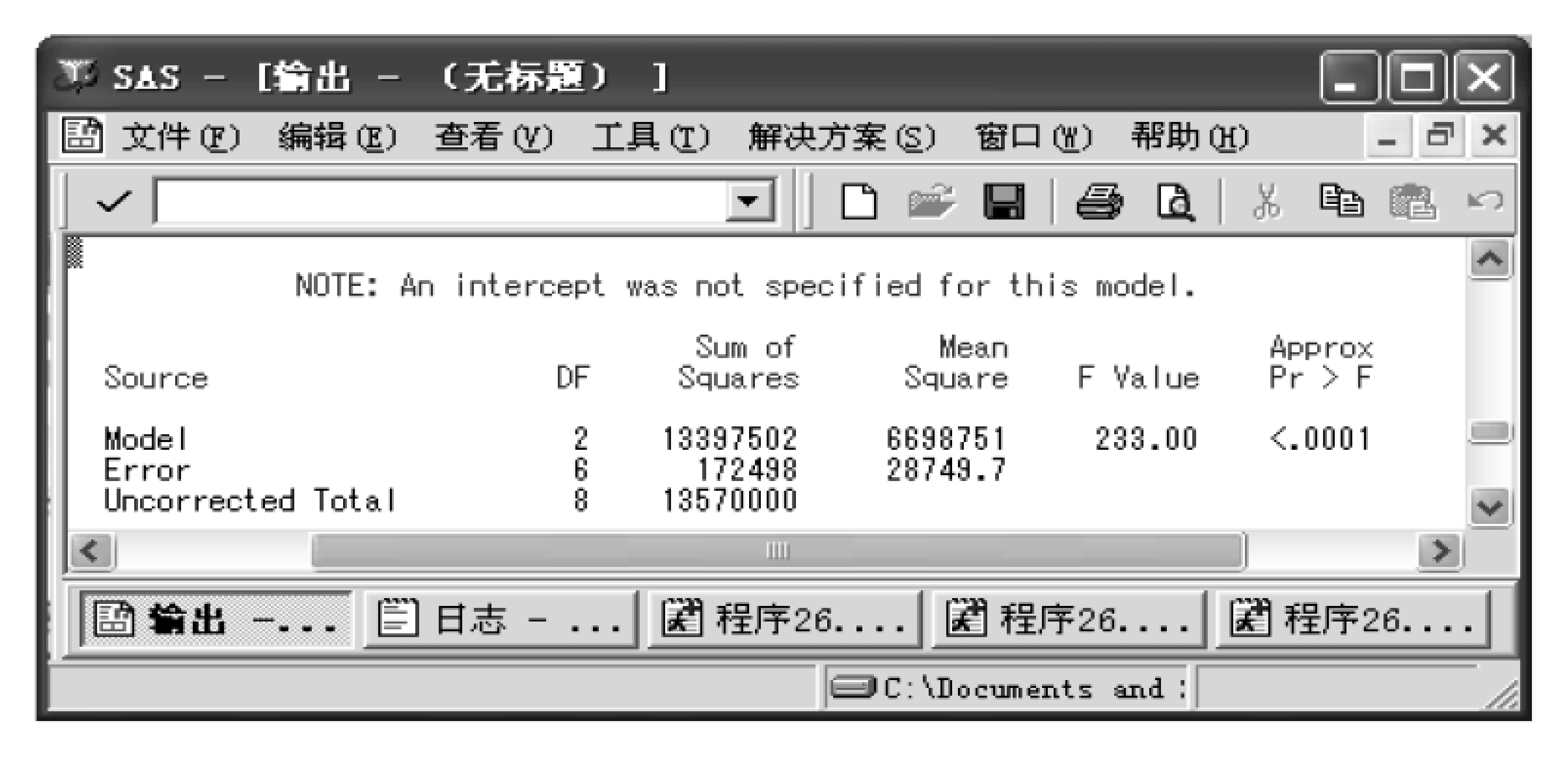


图 15.14　被解释的方差

Error Sum of Squares (Residual SS)：未被解释的残差平方和(其值为 172498)。

Uncorrected Total Sum of Squares：因变量的总平方和(其值为 13570000)。

Corrected Total Sum of Squares：偏离均值的平方和(其值为 2288750，但 SAS 9 不显示)。

R squared＝1－Residual SS/Corrected SS ＝1－172498÷2288750≈0.92：确定系数，或称判定系数为 0.92。此值表示被模型解释的方差占总方差的 92%，模型合格。

(2) 置信度分析。对图 15.15 的统计解释如下。

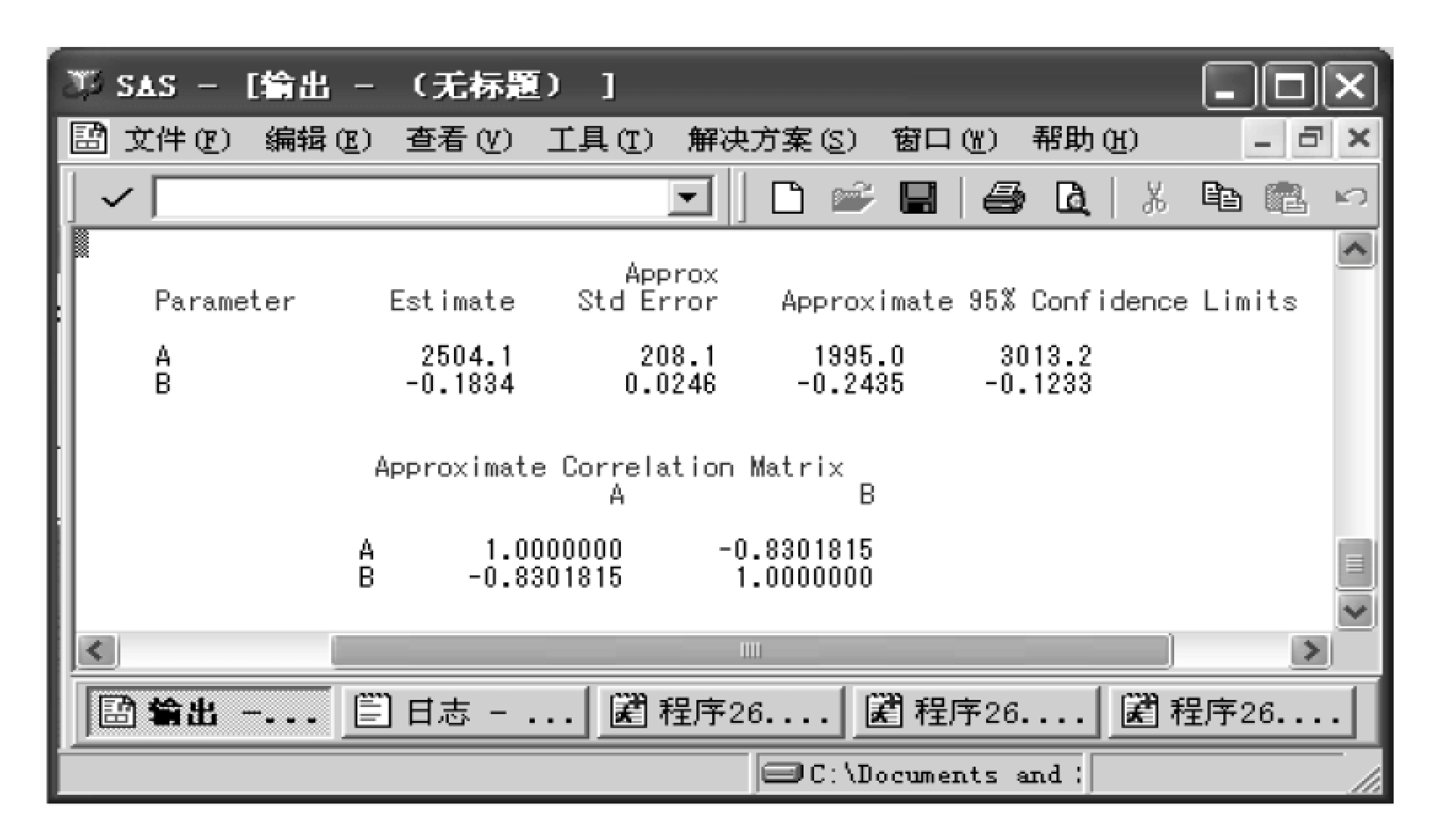


图 15.15　系数的估计值

Parameter：参数，有 A、B 参数。

Estimate：参数的估计值。如 B 参数的估计值为 －0.1834，A 参数的估计值为 2504.1。

Approximate 95% Confidence Limits：逼近的 95%置信度。

左栏(Lower)：95%置信度的下限(如 A 的 95%置信度的下限为 1995.0)。

右栏(Upper)：95%置信度的上限(如 A 的 95%置信度的上限为 3013.2)。

注意：由于本例 95%置信区间的上下限不经过坐标原点，所以可拒绝系数为 0 的原假设。

(3) 逼近的相关矩阵。

从图 15.15 逼近的相关矩阵可看出,参数 A 与 B 的相关系数为 $|-0.8301815|$,很大。所以该模型很可能"超参数"(Over Parameterized)。但出现超参数时不一定意味着模型完全不拟合数据,因为本例的 Cases 太少,可能导致不足以估计全部参数。

15.6.3 指数曲线的预测

最后,获得指数曲线的预测公式如下:

$$\text{decrease} = a * \text{EXP}(b * month) = 2504.1 * \text{EXP}(-0.1834 * month) \quad (15.12)$$

习 题 15

1. SAS 中的 NLIN 过程的迭代法主要有哪几种?
2. 拟合 Logistic 曲线回归的数学表达式是什么?(提示:见第 15 章 15.2 节)
3. 负指数生长曲线的数学模型是什么?(提示:第 15 章 15.4 节)
4. 指数生长曲线的数学模型是什么?(提示:见 15.6 节)

第 16 章 chapter 16

用 Logistic 过程做逻辑斯蒂克回归

本章介绍 Logistic Regression 过程的应用，这是对因变量是二分变量(0 与 1 编码)，自变量是区间(定距)以上变量的回归分析。

当因变量只有两个值(0 与 1)时，若采用多元线性回归，则预测值不能落入 0～1 区间。若采用线性判别分析，虽然允许对自变量直接分组预测，但是自变量正态分布的假设，以及两组协方差相等的假设，需要最佳的预测规则。因此，当因变量只有 0 与 1 这两个值时，应该采用 Logistic Regression 模型估算出一个事件发生的概率。这种模型所需的假设，比判别分析所需的假设简要得多。

16.1 逻辑斯蒂克回归模型

在逻辑斯蒂克(Logistic)回归中，可直接计算一个事件发生的概率。

(1) 对于只有一个自变量的 Logistic Regression 回归模型，可以写成：

$$\mathrm{Prob(event)} = \mathrm{e}^{(B_0+B_1\times X_1)}/(1+\mathrm{e}^{(B_0+B_1\times X_1)}) \tag{16.1}$$

或

$$\mathrm{Prob(event)} = 1/(1+\mathrm{e}^{-(B_0+B_1\times X_1)}) \tag{16.2}$$

式(16.2)中

B_0：回归截距。

B_1：是从数据中计算出的回归系数。

X_1：是自变量。

e：是自然对数的底，e≈2.178。

(2) 对于多个自变量的 Logistic Regression 模型，可以写成式(16.3)。

$$\mathrm{Prob(event)} = \mathrm{e}^{z}/(1+\mathrm{e}^{z}) \tag{16.3}$$

或

$$\mathrm{Prob(event)} = 1/(1+\mathrm{e}^{-z}) \tag{16.4}$$

式(16.4)中，Z 是线性结合模型，即

$$Z = B_0 + B_1X_1 + B_2X_2 + \cdots + BpXp \tag{16.5}$$

(3) 事件未发生的概率，可写成：

$$\mathrm{Prob(no\ event)} = 1 - \mathrm{prob(event)} \tag{16.6}$$

(4) Logistic Regression 的曲线图。

如果能画出 Logistic Regression 的曲线图,从图中便可看到 Logistic 回归曲线呈现 S 型。它非常类似于正态分布的累积概率曲线。但不管 Z 值如何,概率(prob)值总在 0~1 之间。

在线性回归模型中,是用最小二乘方来估算模型的参数。所谓最小二乘方,是指因变量的实际观察值与期望值之间的距离的最小平方和。

在 Logistic 回归模型中,是用最大似然度法(maximum-likelihood method)估计出模型的系数,即这种系数是最接近于观察结果。而且,因为这种 Logistic 回归是非线性的,因此,估算系数时需要采用迭代计算。

16.2 Logistic 回归过程对数据的要求

1980 年,美国科学家 Brown 对 53 名前列腺癌症患者进行预测和研究,他报告了每位患者的年龄(age)、血清磷化物实验结果(如果癌细胞扩散,其值升高),以及疾病的阶段、X 射线的结果。报告的每一项,就是一个变量。然后根据这些变量的数据,预测出淋巴结癌细胞是否呈现阳性。

本章用相似的方法,调查了 350 名大学生,研究恋爱与否(变量 V=0 未恋爱,V=1 已恋爱),与年龄(age)、Location(地区)、各科平均成绩(Score)、性别(sex)以及身高(height)等变量的关系。其中 Location 和 sex 编码为(0~1)两个水平。例如 Location=0 为来自农村的学生,Location=1 为来自城市的学生。sex=0 为女生,sex=1 为男生。详见表 16.1 的编码和程序 16.1。

1. 变量

见表 16.1。

表 16.1 变量定义

V 恋爱	Age 年龄	Location 地区	S 各科平均分	Sex 性别	Height 身高
0:未恋爱 1:已恋爱	单位:岁	0:农村 1:城市	单位:分	0:女 1:男	单位:厘米

2. 数据定义

见程序 16.1。

程序 16.1:SAS 的数据定义。

```
DATA bz96_98;
  INPUT age location vf vm s1 s2 height
    weight like1 like2 like3 v sex will ;
  score= s1/s2 * 100;
  CARDS;                  /* 下面是 19行数据 * /
```

```
20    1 4 2 582    750    168    108    2 4 6 0 1 4
19    1 4 4 502    750    160    98     2 3 7 0 0 5
21    0 2 2 361    750    175    126    1 5 4 0 1 5
21    1 4 1 561    750    170    112    1 4 5 0 1 4
21    1 4 1 558    900    158    110    1 2 3 0 0 7
20    0 1 2 465    750    168    128    5 0 0 0 1 6
19    1 5 5 549    750    0      0      2 1 7 0 0 1
22    1 1 1 382    750    156    110    2 3 7 0 0 7
21    0 2 2 595    750    166    112    2 4 6 1 1 5
20    1 3 4 490    750    158    98     3 0 0 0 0 1
20    0 1 1 409    650    178    140    3 4 6 1 1 7
20    1 6 6 436    750    164    128    2 4 6 0 1 5
20    1 3 3 421    750    168    84     1 2 6 1 1 7
20    0 1 6 615    900    165    106    2 4 7 0 0 5
22    1 4 4 450    750    170    160    2 7 0 0 1 1
21    1 4 4 0      0      0      0      0 0 0 0 1 7
23    1 4 1 482    750    168    106    1 2 7 0 0 5
20    1 1 1 475    750    170    120    3 4 6 1 1 0
18    0 4 1 0      0      160    106    7 0 0 0 0 1
;
RUN;
```

首先应在图 16.1 的编辑器中编辑程序 16.1，然后可按照 16.3 节进行统计。

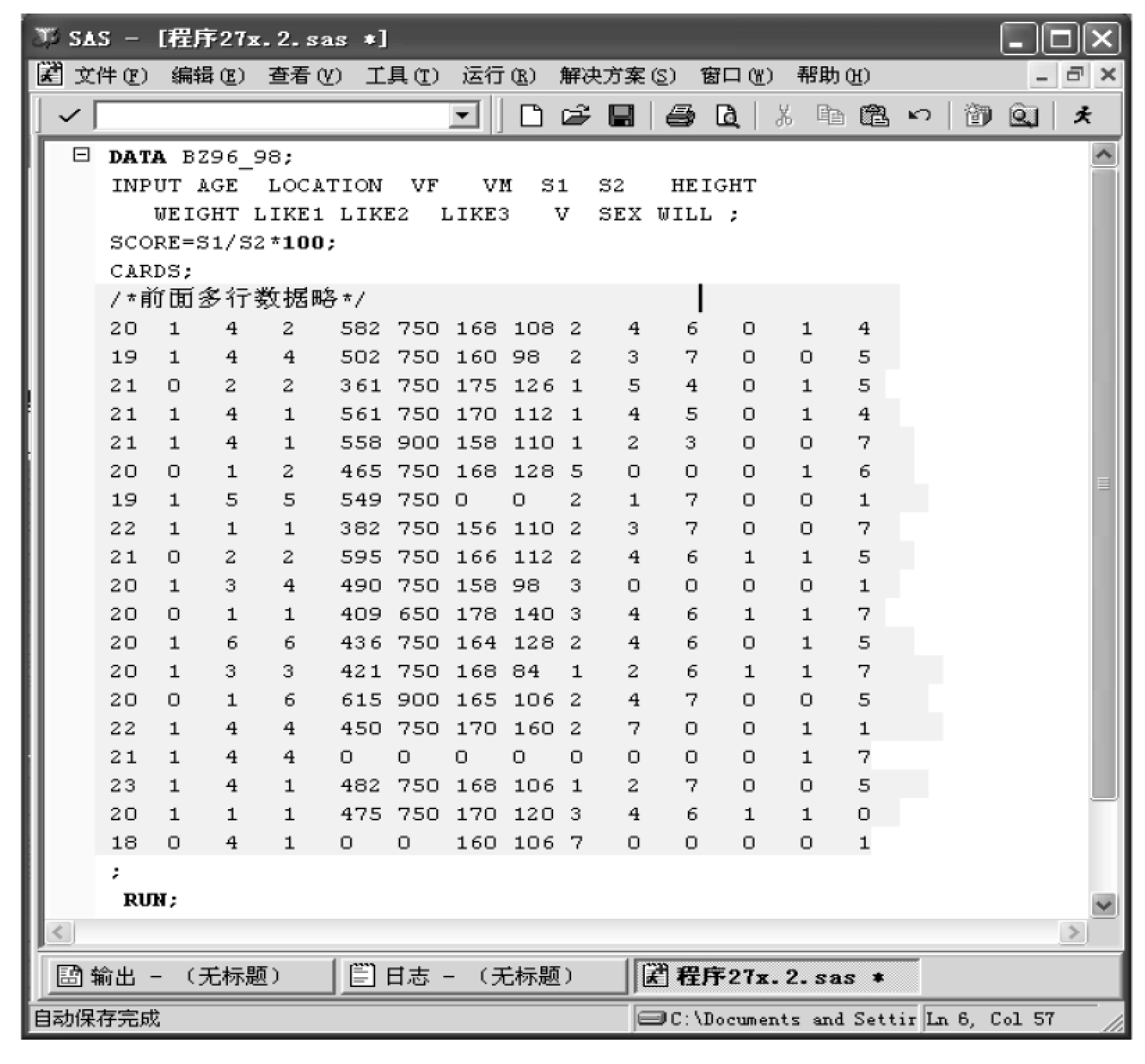


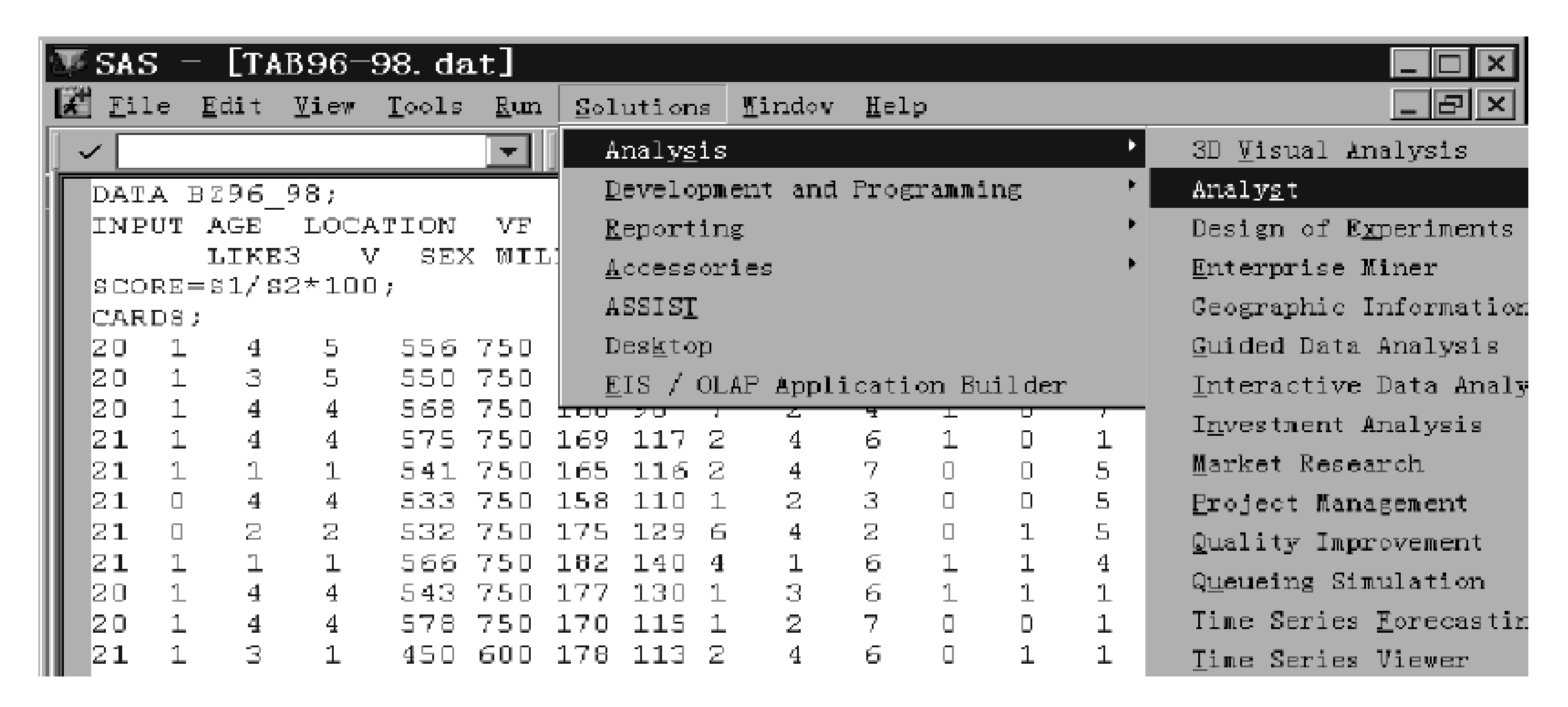
图 16.1　程序 16.1 的数据编辑

16.3 用“分析家”对话框做 Logistic 回归

1. 选择“分析家”的操作步骤

(1) 选择图 16.1 主菜单中的“运行”→“提交”命令，产生 SAS 数据集 Work. BZ96_98。

(2) 选择 SAS 主菜单中的“解决方案”→“分析”命令，鼠标指针移到图 16.2 中带有阴影标记的“分析家”命令上。



(a) SAS 8e的命令位置



(b) SAS 9的命令位置

图 16.2 分析家的命令位置

(3) 选择“分析家”→“文件”→“按 SAS 名称打开”命令后进入图 16.3，并且查找出“逻辑库”中的“Work. Bz96_98”数据集。

(4) 选择“Bz96_98”后单击“确定”(或 OK)按钮，显示 Work. Bz96_98 数据集的内容，见图 16.4。

(5) 选择“统计”→“回归”命令 SAS 显示图 16.5。

(6) 选择 Logistic 命令进入图 16.6 并设置变量。

说明：age 等连续变量必须放在图 16.6 的 Quantitative 框内，location 等标称变量作



图 16.3　查找 Work.Bz96_98 数据集文件

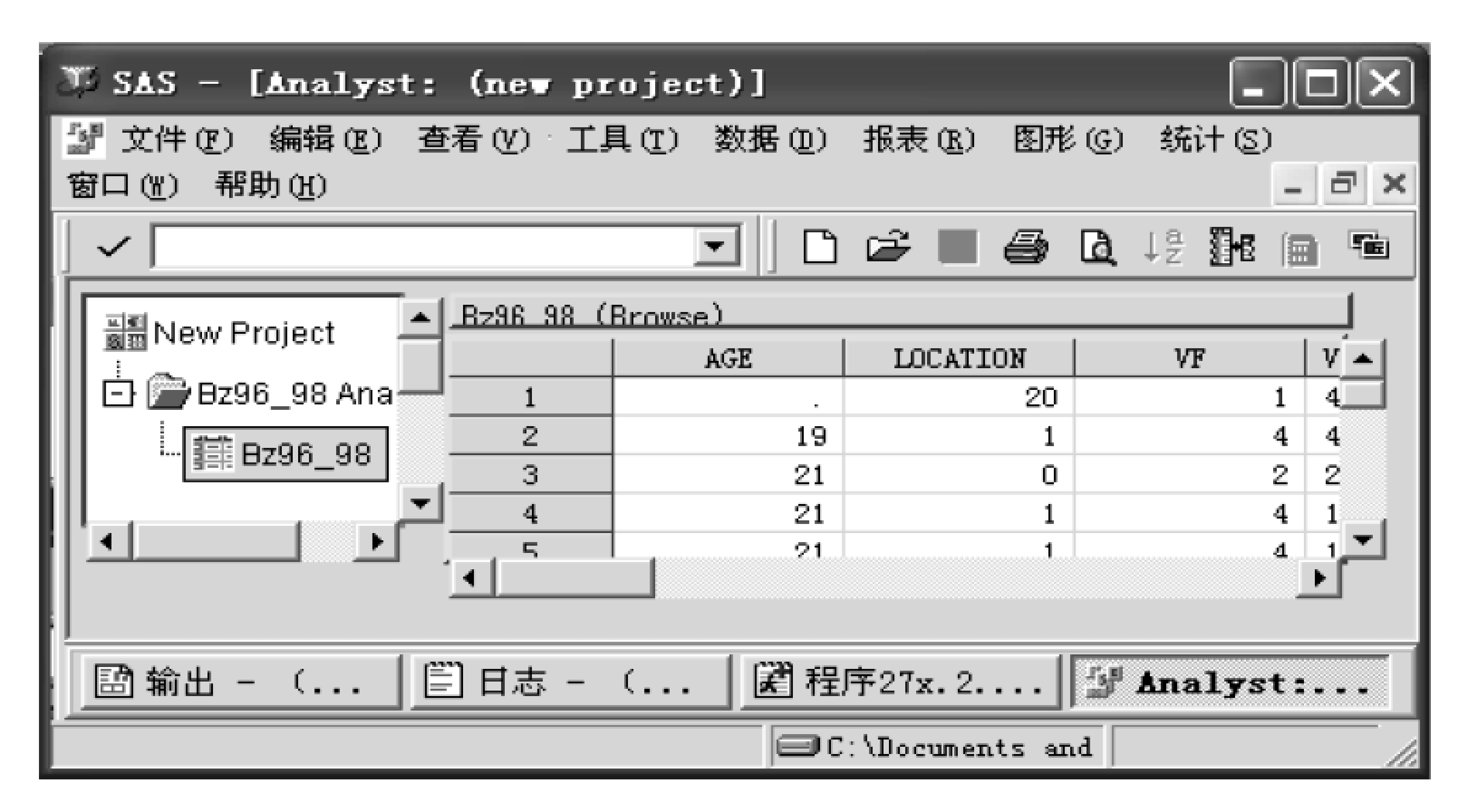


图 16.4　Work.Bz96_98 文件的内容(部分)



图 16.5　Regression 的菜单位置

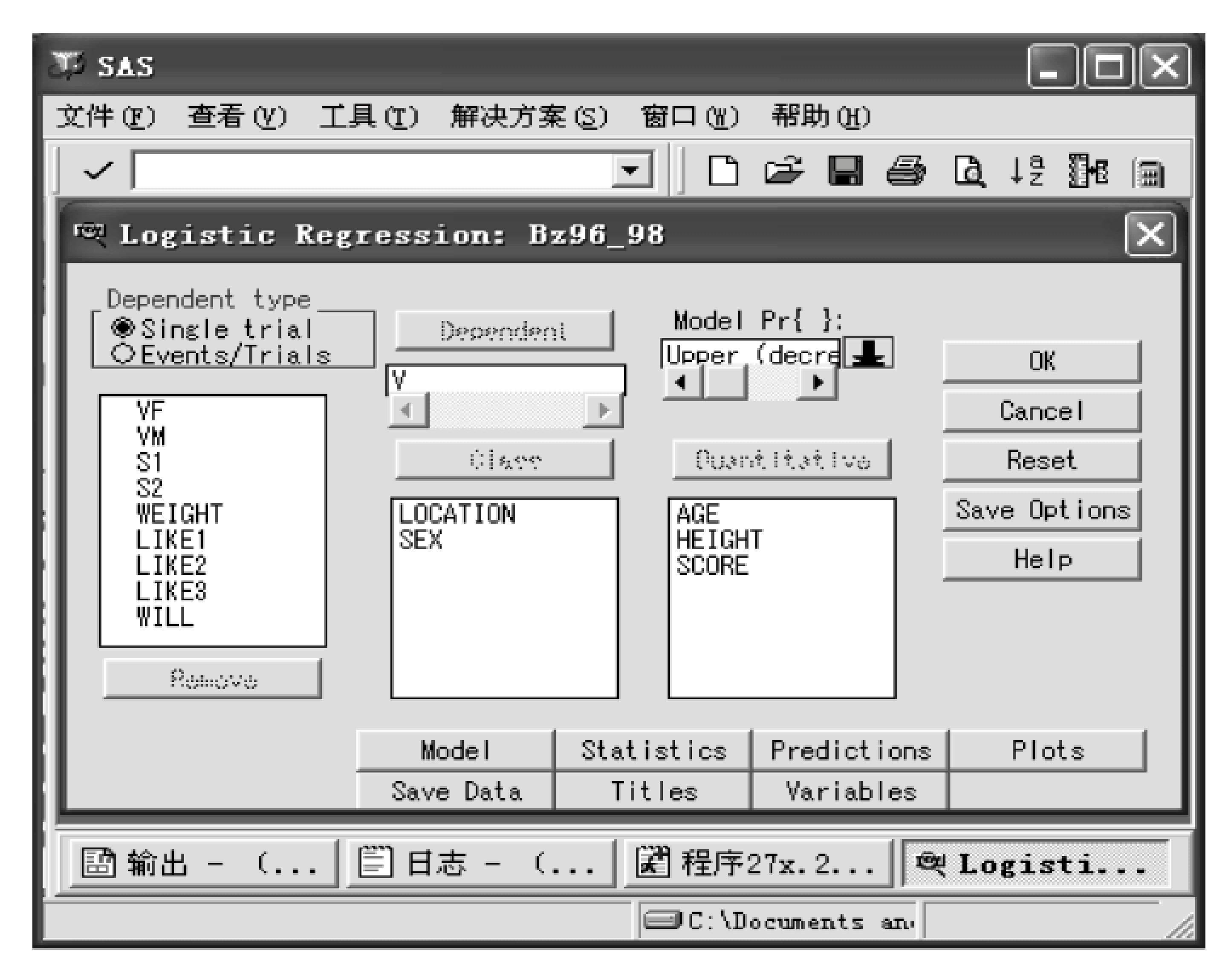


图 16.6　选择 Logistic 回归的因变量 v(恋爱)和自变量 age 等

为分类变量只能放在 Class 框内。本例将计算 V＝1(事件已经发生)的概率。

(7) 单击 Model 按钮进入图 16.7 中,再单击 Standard Models 按钮,选择默认的主效应模型。

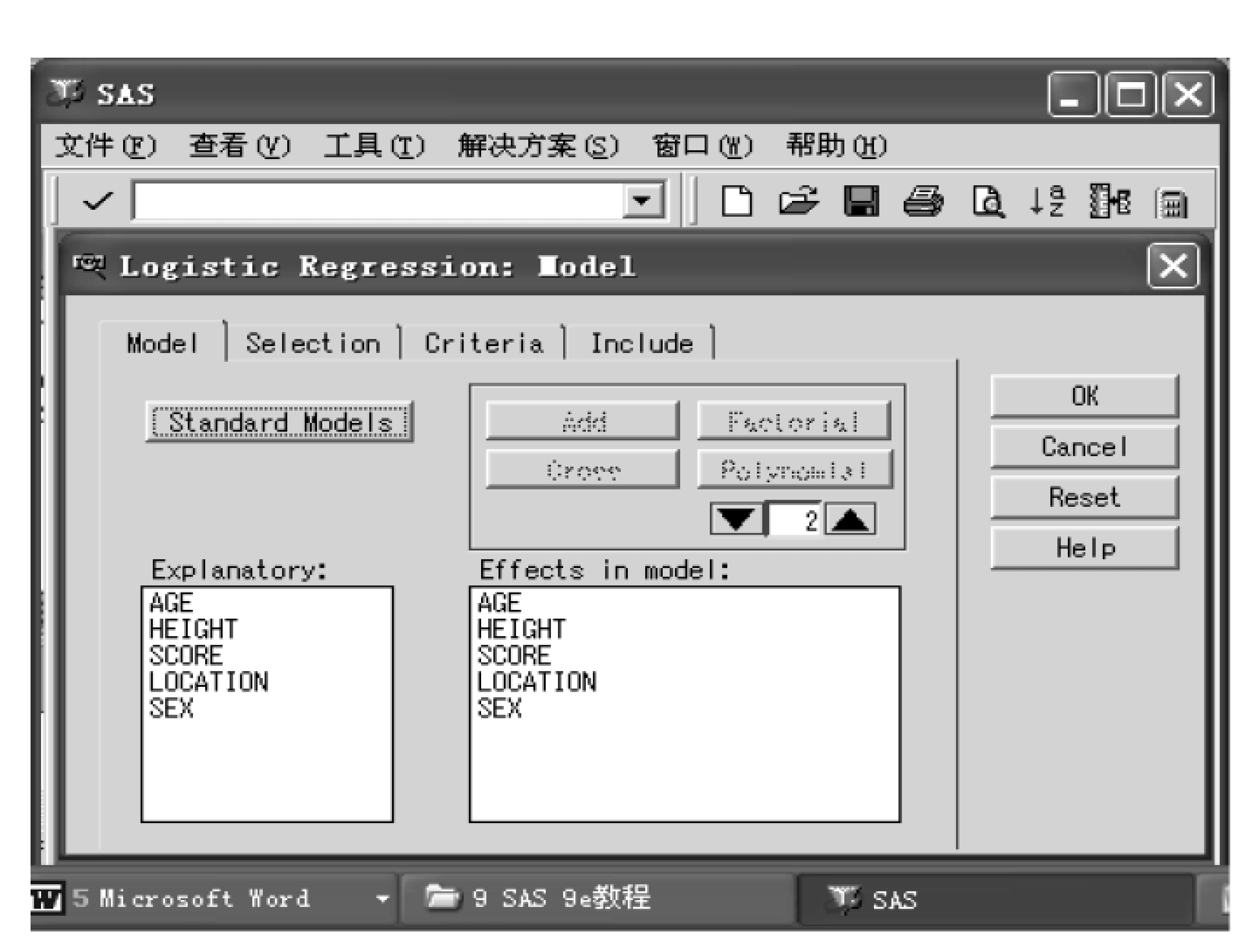


图 16.7　选择主效应(默认)模型

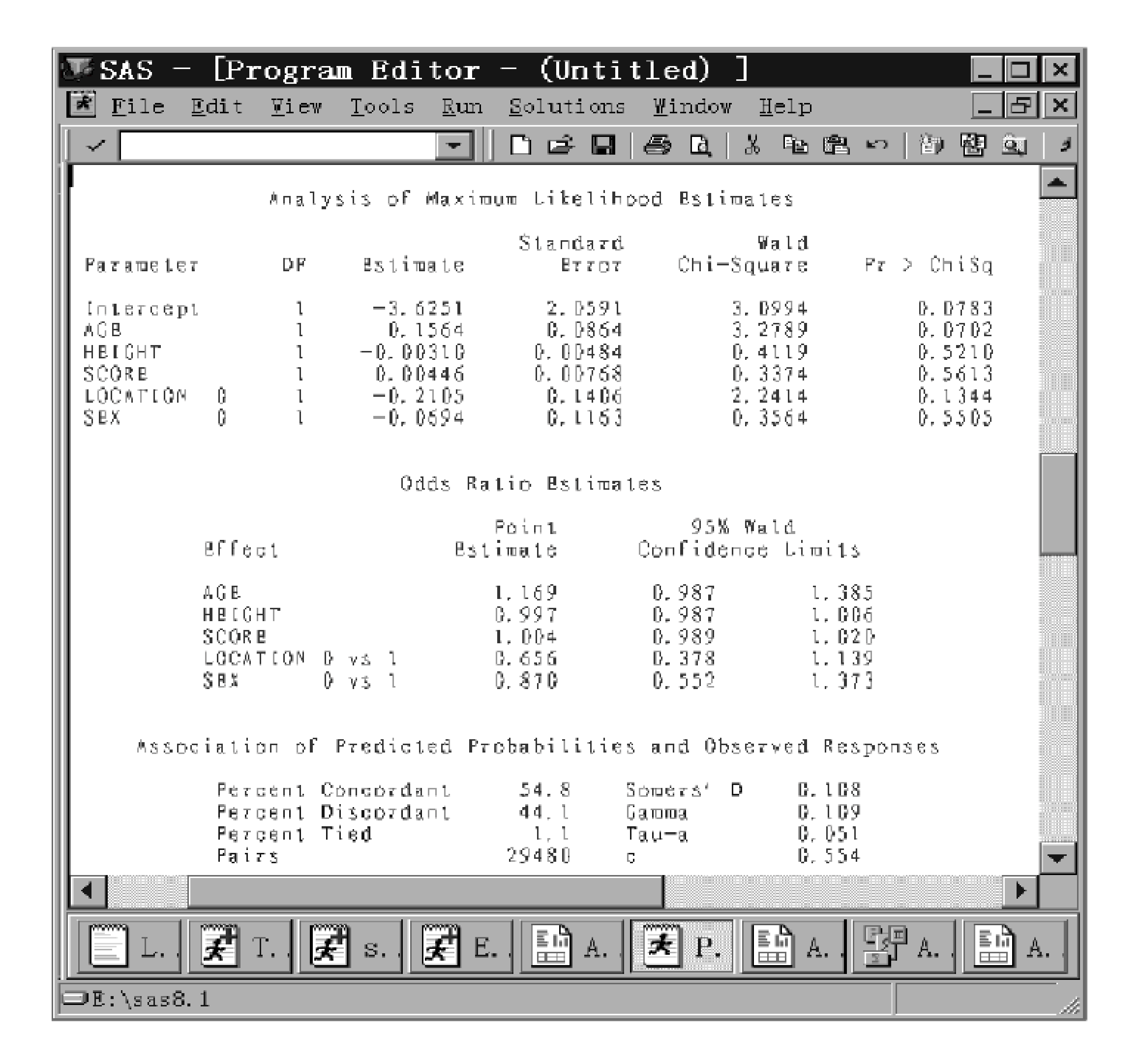
(8) 单击 OK 按钮产生图 16.8。

(9) 再单击图 16.6 中的 Model 按钮,进入图 16.9 并选择主效应和部分的二次项效应。

(10) 单击 OK 按钮产生图 16.10。

2. 对图 16.8 和图 16.10 的结果比较

图 16.8 是模型只有主效应时的输出结果,Pr>Chisq 一栏的显著性水平都大于 α 值



SAS - [Program Editor - (Untitled)]

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-3.6251	2.0591	3.0994	0.0783
AGE		1	0.1564	0.0864	3.2789	0.0702
HEIGHT		1	-0.00310	0.00484	0.4119	0.5210
SCORE		1	0.00446	0.00768	0.3374	0.5613
LOCATION	0	1	-0.2105	0.1406	2.2414	0.1344
SEX	0	1	-0.0694	0.1163	0.3564	0.5505

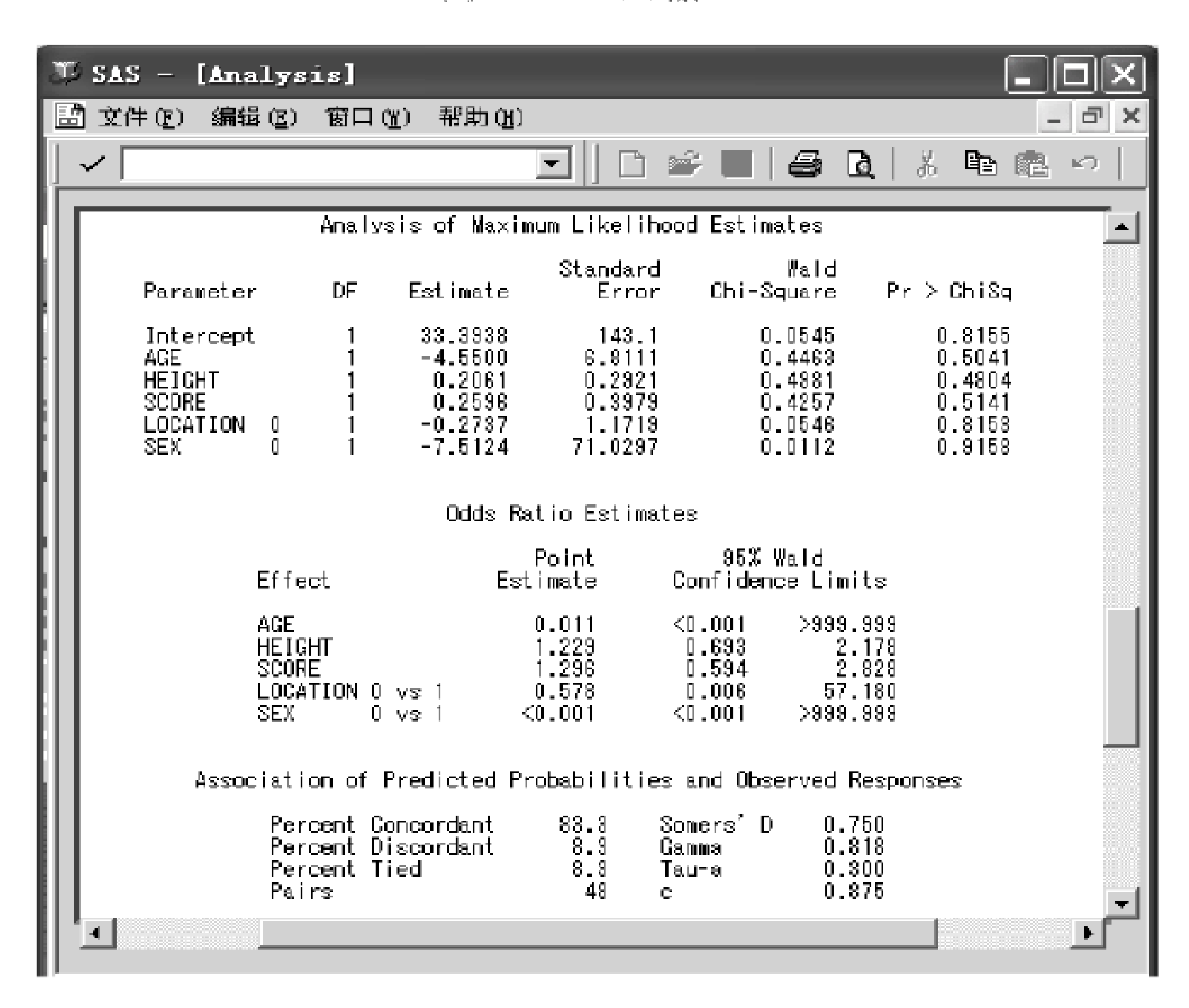
Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
AGE	1.169	0.987	1.385
HEIGHT	0.997	0.987	1.006
SCORE	1.004	0.989	1.020
LOCATION 0 vs 1	0.656	0.378	1.139
SEX 0 vs 1	0.870	0.552	1.373

Association of Predicted Probabilities and Observed Responses

Percent Concordant	54.8	Somers' D	0.108
Percent Discordant	44.1	Gamma	0.109
Percent Tied	1.1	Tau-a	0.051
Pairs	29480	c	0.554

(a) SAS 8e的输出



SAS - [Analysis]

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	33.3938	143.1	0.0545	0.8155
AGE		1	-4.5500	6.8111	0.4463	0.5041
HEIGHT		1	0.2061	0.2821	0.4881	0.4804
SCORE		1	0.2598	0.3979	0.4257	0.5141
LOCATION	0	1	-0.2737	1.1719	0.0546	0.8153
SEX	0	1	-7.5124	71.0297	0.0112	0.9158

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
AGE	0.011	<0.001	>999.999
HEIGHT	1.229	0.693	2.178
SCORE	1.296	0.594	2.828
LOCATION 0 vs 1	0.578	0.006	57.180
SEX 0 vs 1	<0.001	<0.001	>999.999

Association of Predicted Probabilities and Observed Responses

Percent Concordant	83.3	Somers' D	0.750
Percent Discordant	8.3	Gamma	0.818
Percent Tied	8.3	Tau-a	0.300
Pairs	48	c	0.875

(b) SAS 9的输出

图 16.8 主效应模型的主要输出部分

0.05，说明截距项和 age 等自变量的回归系数都不显著。

而且，图 16.10 中增加了部分二次项效应之后，Pr>Chisq 一栏的显著性水平（除了 SEX 变量和截距外）仍然大于 α 值 0.05，表明其他回归系数仍然不显著。请读者增加 200 个数据然后重新运行程序 16.1 再做一遍 Logistic 回归，观察图 16.10 中 sex 变量的

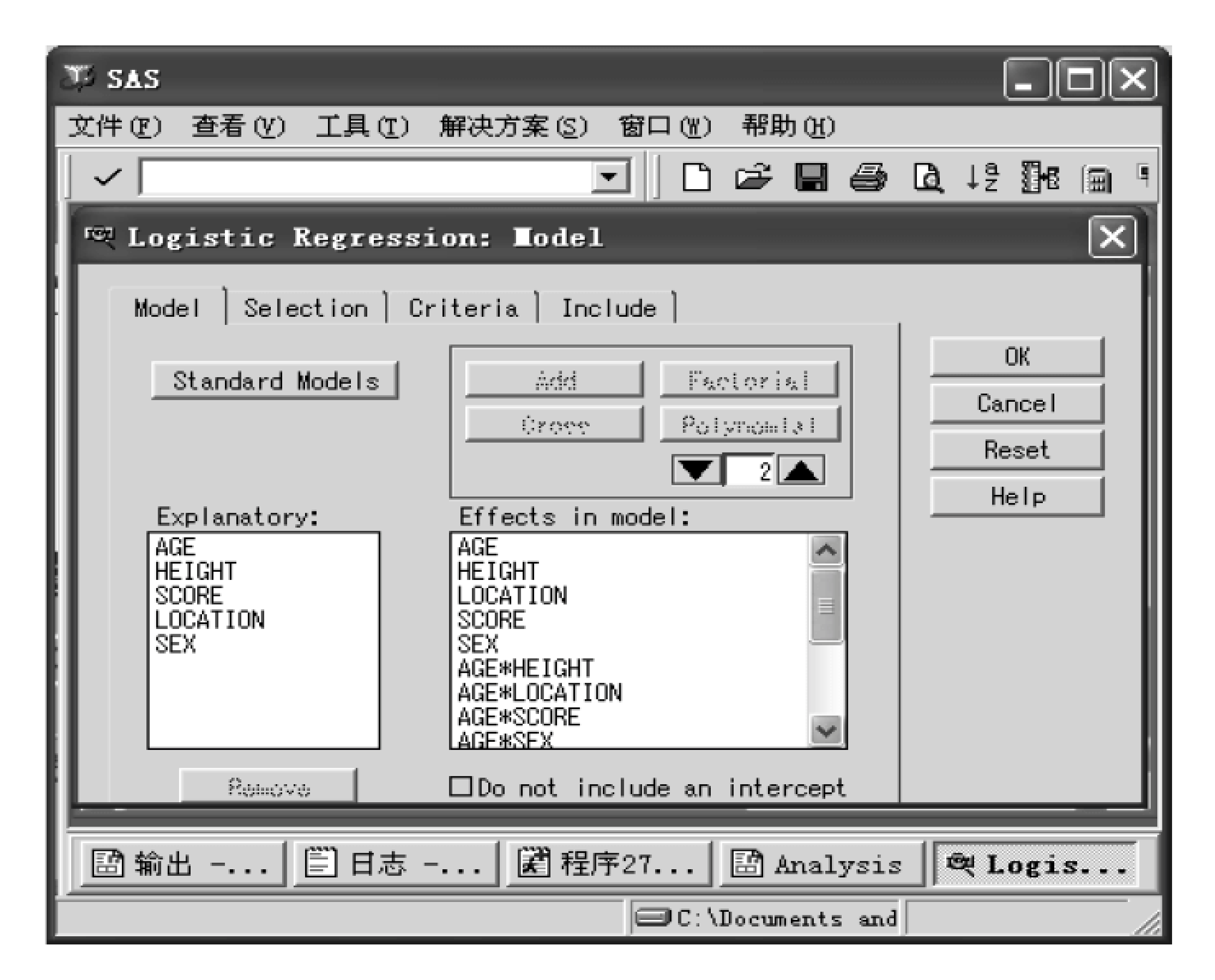


图 16.9　主效应和部分的二次项效应

回归系数的 Pr>Chisq 值,若是小于 α 值 0.05,则由原来的不显著变为显著的了。如果从图 16.10 中能看到性别与年龄之间具有交互效应、性别与地区之间基本上也存在交互效应,就说明恋爱发生与否,还取决于性别与年龄之间的关系。

注意:图 16.8 结果类似于后面的图 16.16,其结果分析很重要,所以在 16.5 节中单独加以详细介绍,如果想先了解结果分析,则跳过 16.4 节而直接阅读 16.5 节的结果分析。

```
                   Analysis of Maximum Likelihood Estimates

                                          Standard        Wald
Parameter            DF    Estimate       Error      Chi-Square    Pr > ChiSq

Intercept             1    -12.1528      5.4948        4.8915        0.0270
AGE                   1      0.0658      0.1072        0.3763        0.5396
HEIGHT                1      0.0582      0.0304        3.6648        0.0556
LOCATION        0     1     -8.8279      4.9164        3.2241        0.0726
SCORE                 1      0.0101     0.00806        1.5749        0.2095
SEX             0     1    -10.1010      3.5329        8.1743        0.0042
AGE*SEX         0     1      0.2973      0.1070        7.7223        0.0055
HEIGHT*LOCATION 0     1      0.0521      0.0294        3.1261        0.0770
HEIGHT*SEX      0     1      0.0254      0.0178        2.0515        0.1521
LOCATION*SEX    0 0   1      0.3763      0.1971        3.6457        0.0562

                          Odds Ratio Estimates

                              Point          95% Wald
          Effect            Estimate      Confidence Limits

          SCORE               1.010        0.994       1.026

                   14:11 Wednesday, June 26, 2002   3

                          The LOGISTIC Procedure

        Association of Predicted Probabilities and Observed Responses

           Percent Concordant     63.9    Somers' D    0.285
           Percent Discordant     35.5    Gamma        0.286
           Percent Tied            0.6    Tau-a        0.134
           Pairs                 29480    c            0.642
```

(a) SAS 8e的输出

图 16.10　主效应和部分二次项效应的输出结果

```
文件(F) 编辑(E) 窗口(W) 帮助(H)

                 Analysis of Maximum Likelihood Estimates

                                        Standard        Wald
Parameter            DF    Estimate       Error    Chi-Square    Pr > ChiSq

Intercept             1      1227.1      296674        0.0000        0.9967
AGE                   1    -85.1574      7177.8        0.0001        0.9928
HEIGHT                1     -4.4699      1884.2        0.0000        0.9981
LOCATION        0     1     -8.6072      7364.5        0.0000        0.9991
SCORE                 1     -9.4036      2728.0        0.0000        0.9972
SEX             0     1      251.4       6549.8        0.0015        0.9694
AGE*HEIGHT            1      0.2468     40.2044        0.0000        0.9951
AGE*LOCATION    0     1      7.1873       617.7        0.0001        0.9907
AGE*SCORE             1      0.3920     35.1034        0.0001        0.9911
AGE*SEX         0     1      4.8833       260.0        0.0004        0.9850
HEIGHT*LOCATION 0     1     -1.2239     40.9286        0.0009        0.9761
HEIGHT*SCORE          1     0.00830     17.4977        0.0000        0.9996
HEIGHT*SEX      0     1     -2.1143       109.3        0.0004        0.9846
SCORE*LOCATION  0     1      1.0159      7.1787        0.0200        0.8875
LOCATION*SEX    0 0   1     -1.2416       108.5        0.0001        0.9909
SCORE*SEX       0     1      0.0183       142.0        0.0000        0.9999
                                                    18:17 Monday, January 12,

                         The LOGISTIC Procedure
The validity of the model fit is questionable.

        Association of Predicted Probabilities and Observed Responses

            Percent Concordant     100.0    Somers' D    1.000
            Percent Discordant       0.0    Gamma        1.000
            Percent Tied             0.0    Tau-a        0.400
            Pairs                     48    c            1.000
```

(b) SAS 9的输出

图 16.10（续）

16.4 用编程法做逻辑斯蒂克回归

本节用 PROC Logistic 编程法进行逻辑斯蒂克回归，即由 PROC Logistic 过程对恋爱(变量 V=0 未恋爱，V=1 已恋爱)与年龄(age)、地区(Location)、学习成绩(Score)、性别(sex)以及身高(height)等 5 个变量的逻辑斯蒂克回归。

1. PROC Logistic 过程所配备的语句

PROC Logistic 过程的语句极其简单，即

```
PROC LOGISTIC ;
MODEL v= age Location Score sex height;
RUN;
```

将上述 3 条语句加入到程序 16.1 的最后，整个程序则成了可以执行的程序，见程序 16.2。

程序 16.2：一个可执行的简单程序(BZ96_98.dat)。

```
DATA BZ96_98;
  INPUT age location vf vm s1 s2 HEIGHT
    WEIGHT LIKE1 LIKE2 LIKE3 V sex WILL ;
  SCORE= s1/s2 * 100;
  CARDS;
/* 下面是 19行数据 * /
 20  1 4  2  582  750  168  108  2 4 6 0 1 4
```

```
19  1 4  4 502  750  160  98   2 3 7 0 0 5
21  0 2  2 361  750  175  126  1 5 4 0 1 5
21  1 4  1 561  750  170  112  1 4 5 0 1 4
21  1 4  1 558  900  158  110  1 2 3 0 0 7
20  0 1  2 465  750  168  128  5 0 0 0 1 6
19  1 5  5 549  750  0    0    2 1 7 0 0 1
22  1 1  1 382  750  156  110  2 3 7 0 0 7
21  0 2  2 595  750  166  112  2 4 6 1 1 5
20  1 3  4 490  750  158  98   3 0 0 0 0 1
20  0 1  1 409  650  178  140  3 4 6 1 1 7
20  1 6  6 436  750  164  128  2 4 6 0 1 5
20  1 3  3 421  750  168  84   1 2 6 1 1 7
20  0 1  6 615  900  165  106  2 4 7 0 0 5
22  1 4  4 450  750  170  160  2 7 0 0 1 1
21  1 4  4 0    0    0    0    0 0 0 0 1 7
23  1 4  1 482  750  168  106  1 2 7 0 0 5
20  1 1  1 475  750  170  120  3 4 6 1 1 0
18  0 4  1 0    0    160  106  7 0 0 0 0 1
;
PROC LOGISTIC;
    MODEL v= age Location Score sex height;
RUN;
```

首先在图 16.11 的程序编辑器中编辑程序 16.2。

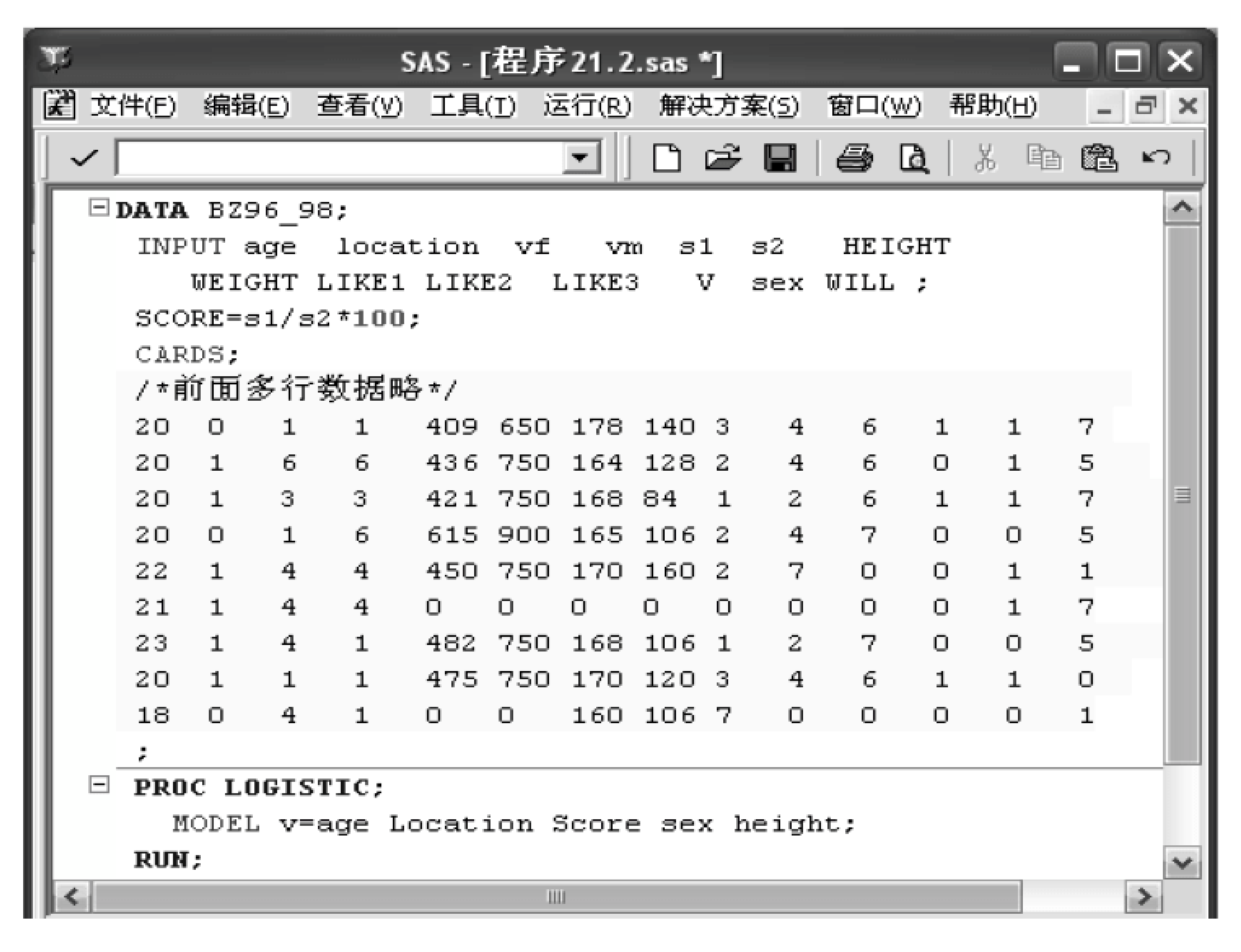


图 16.11 编辑程序 16.2

接着按 F8 键(或从图 16.12 中选择“运行”→“提交”命令)产生数据集 Work. Bz96_98 并产生图 16.13 所示的输出结果。

SAS - [程序27x.2.sas]
文件(F) 编辑(E) 查看(V) 工具(T) 运行(R) 解决方案(S) 窗口(W) 帮助(H)
提交(S)
重新调用上一次提交(R)
提交第一行(T)
提交 N 行(N)...
登录(I)...
远程提交(B)...
远程获取(O)...
远程显示(D)...
注销(F)...

20	1	4	2	582	750	4
19	1	4	4	502	750	5
21	0	2	2	361	750	5
21	1	4	1	561	750	4
21	1	4	1	558	900	7
20	0	1	2	465	750	6
19	1	5	5	549	750	1
22	1	1	1	382	750	7

输出 - (无标题)　程序27x.2.sas
提交语句进行处理　C:\Documents and Setti　Ln 23, Col 59

图 16.12　编辑运行程序 16.2

2. 输出结果

输出结果如图 16.13～图 16.18 所示。

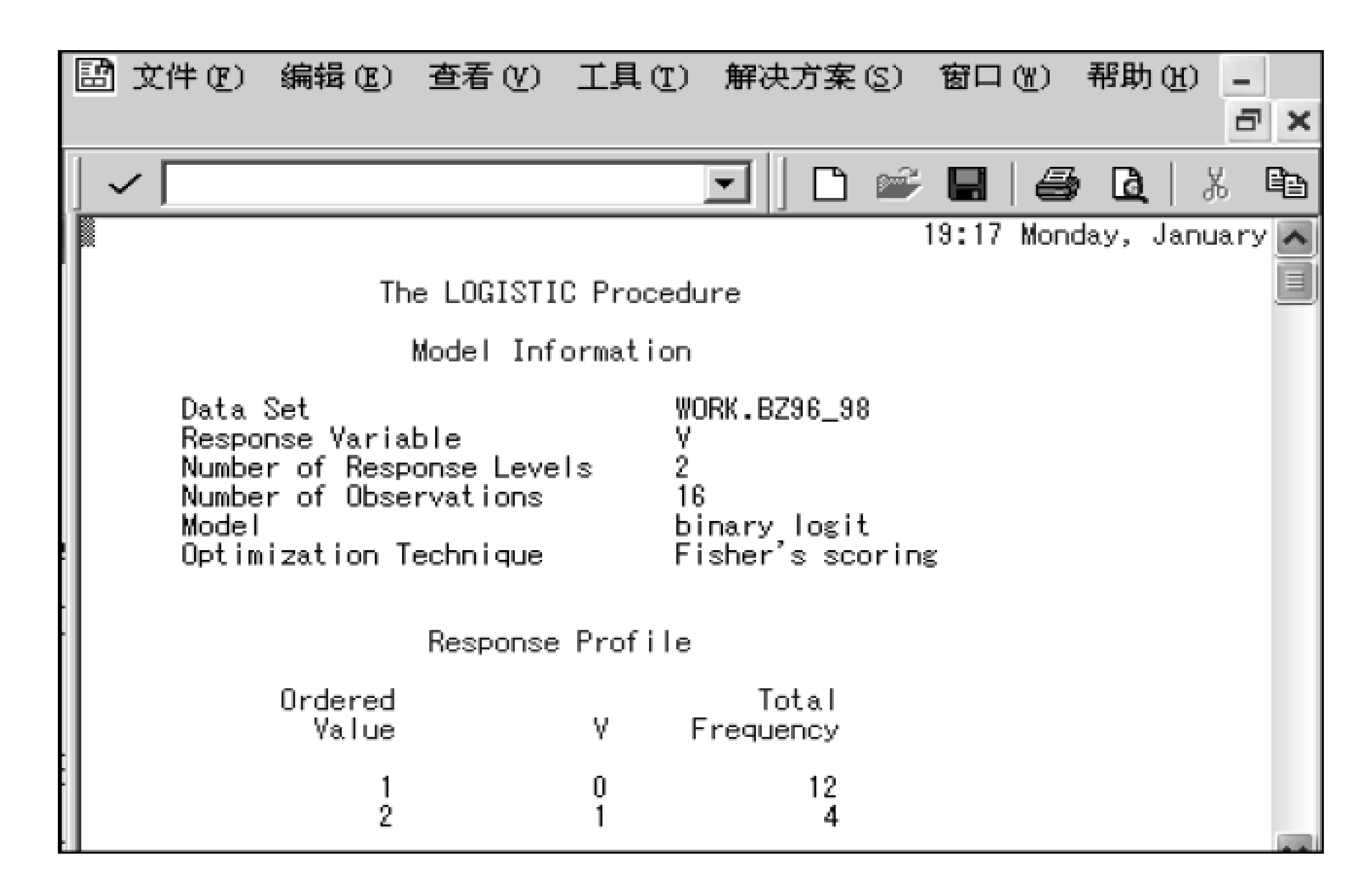


图 16.13　模型中观察值分布情况

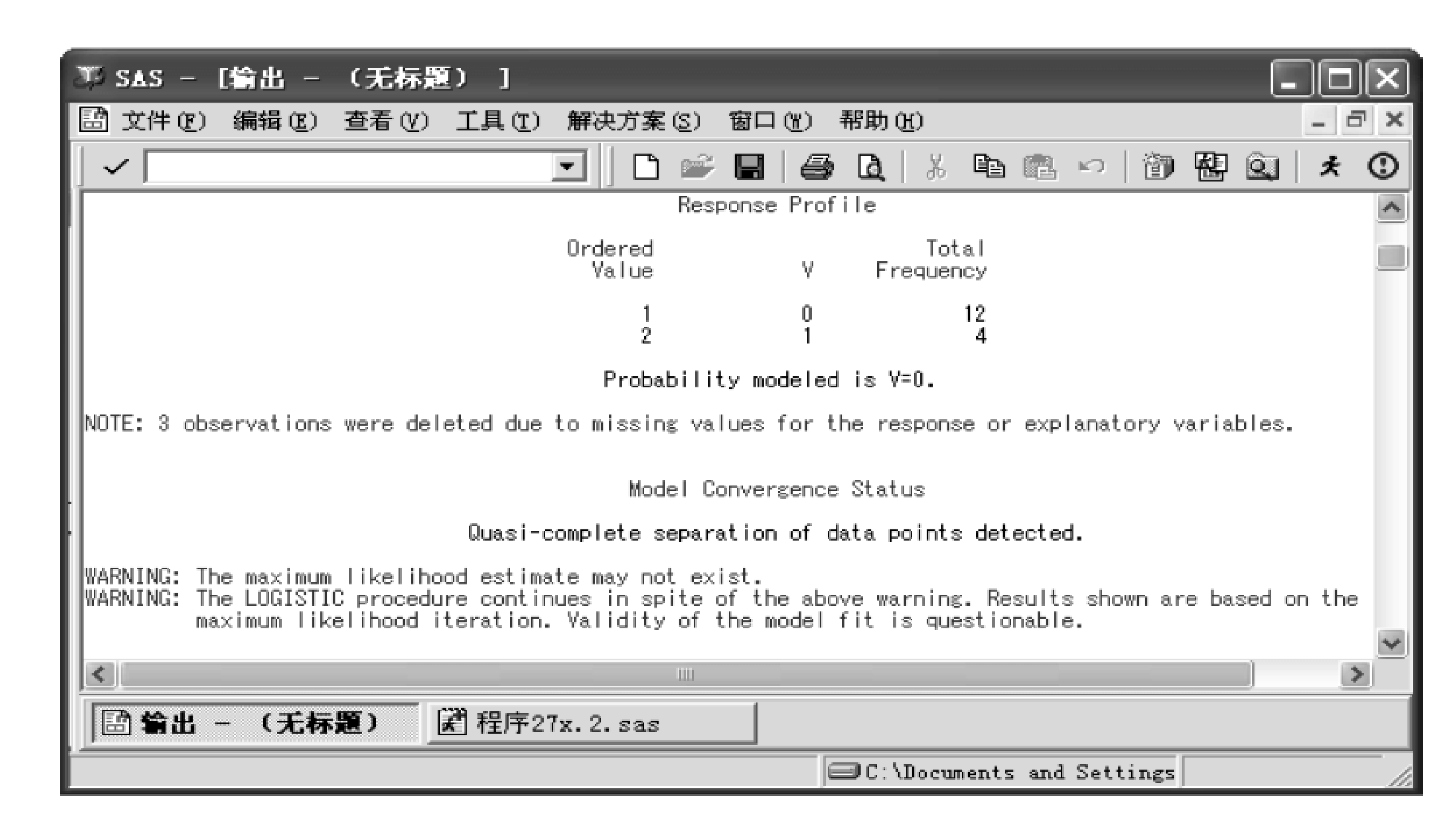
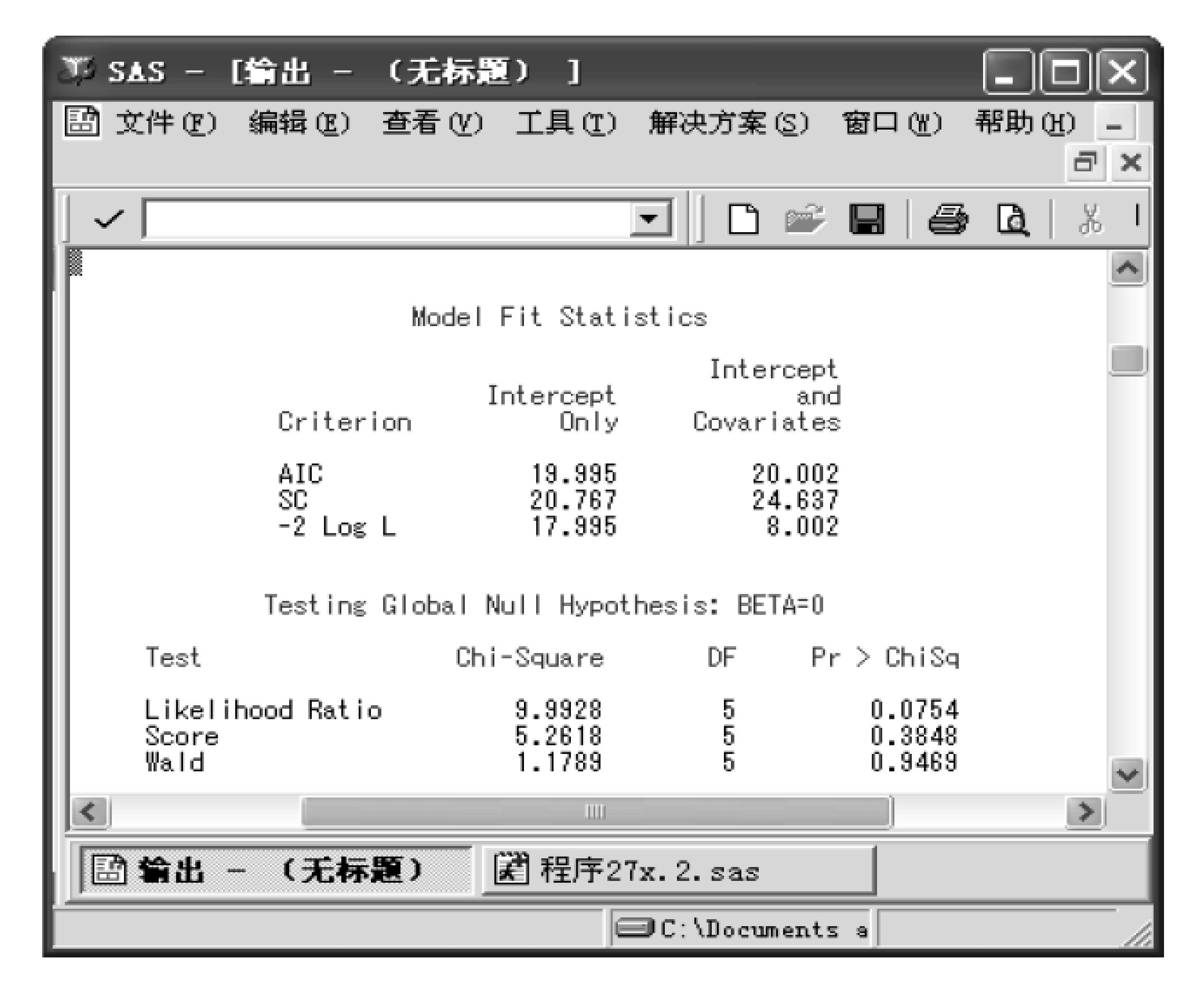


图 16.14　因变量 V 的编码方法及频数分布



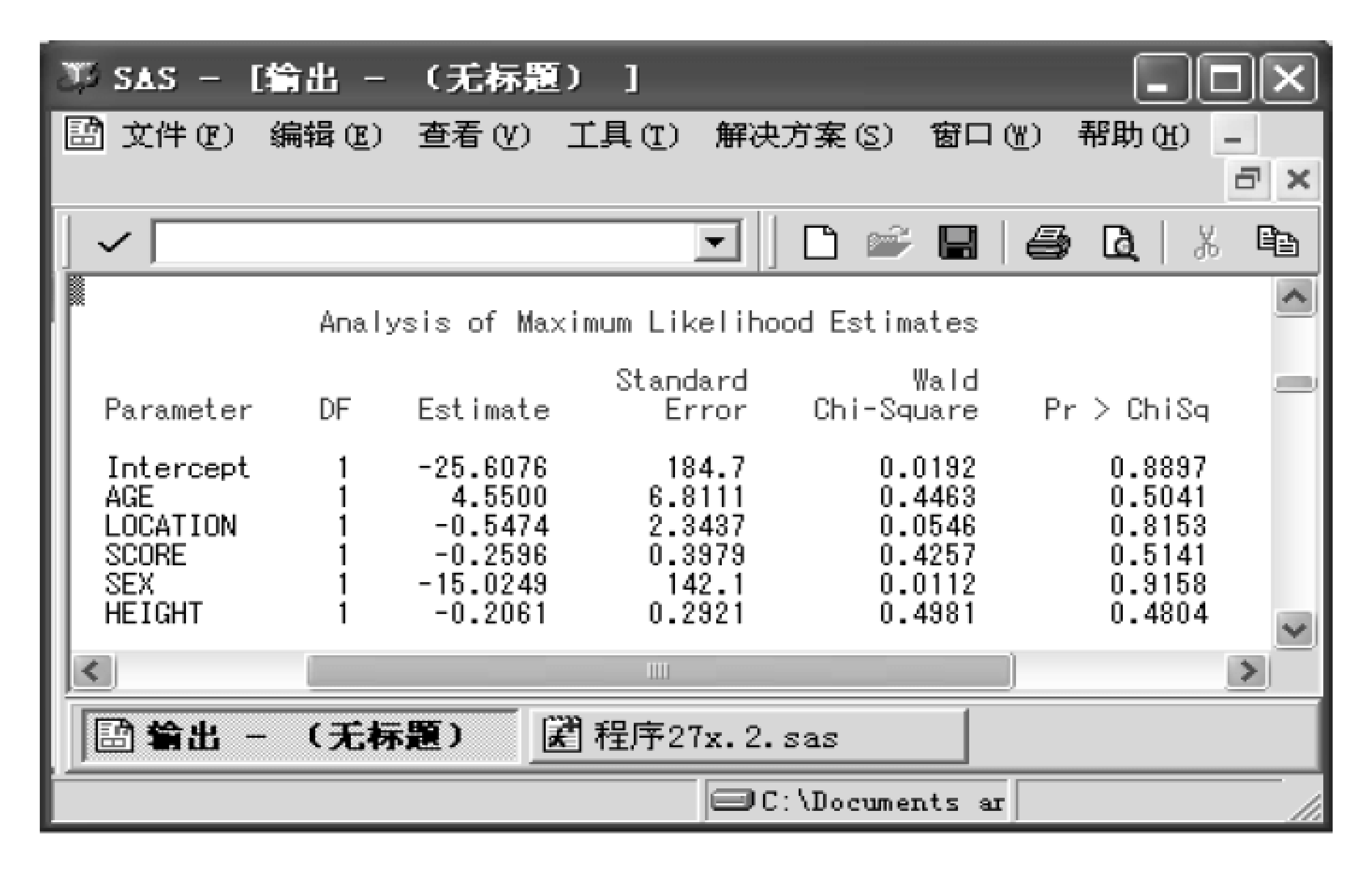
Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	19.995	20.002
SC	20.767	24.637
-2 Log L	17.995	8.002

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	9.9928	5	0.0754
Score	5.2618	5	0.3848
Wald	1.1789	5	0.9469

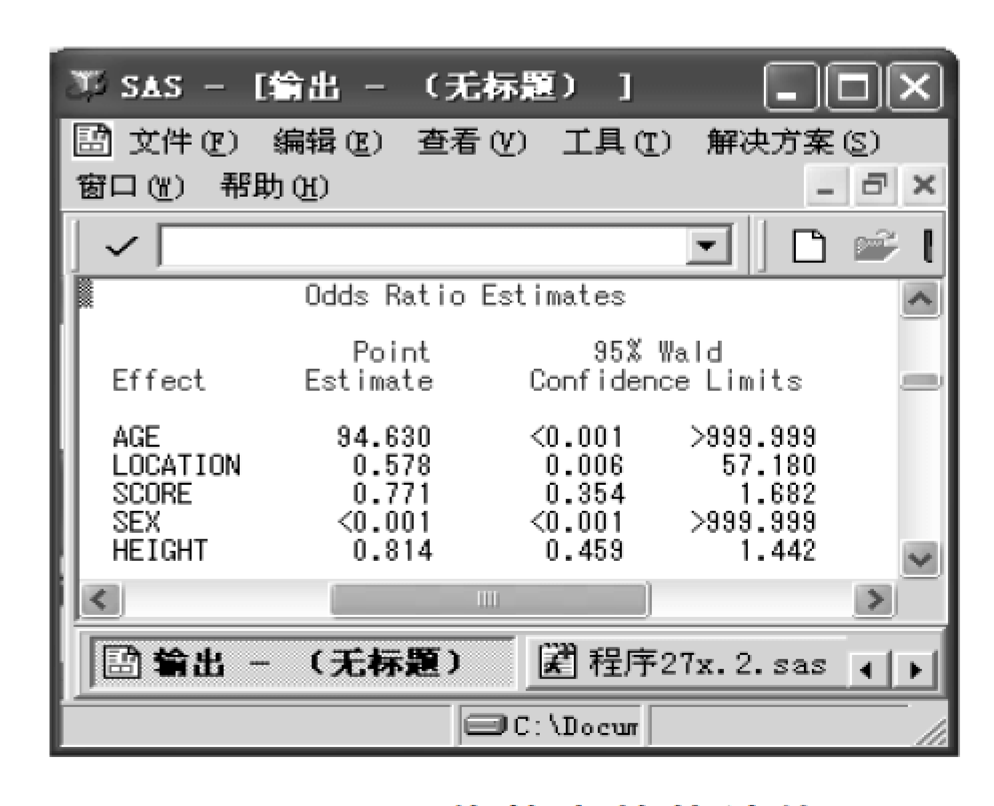
图 16.15　模型拟合度统计量



Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-25.6076	184.7	0.0192	0.8897
AGE	1	4.5500	6.8111	0.4463	0.5041
LOCATION	1	-0.5474	2.3437	0.0546	0.8153
SCORE	1	-0.2596	0.3979	0.4257	0.5141
SEX	1	-15.0249	142.1	0.0112	0.9158
HEIGHT	1	-0.2061	0.2921	0.4981	0.4804

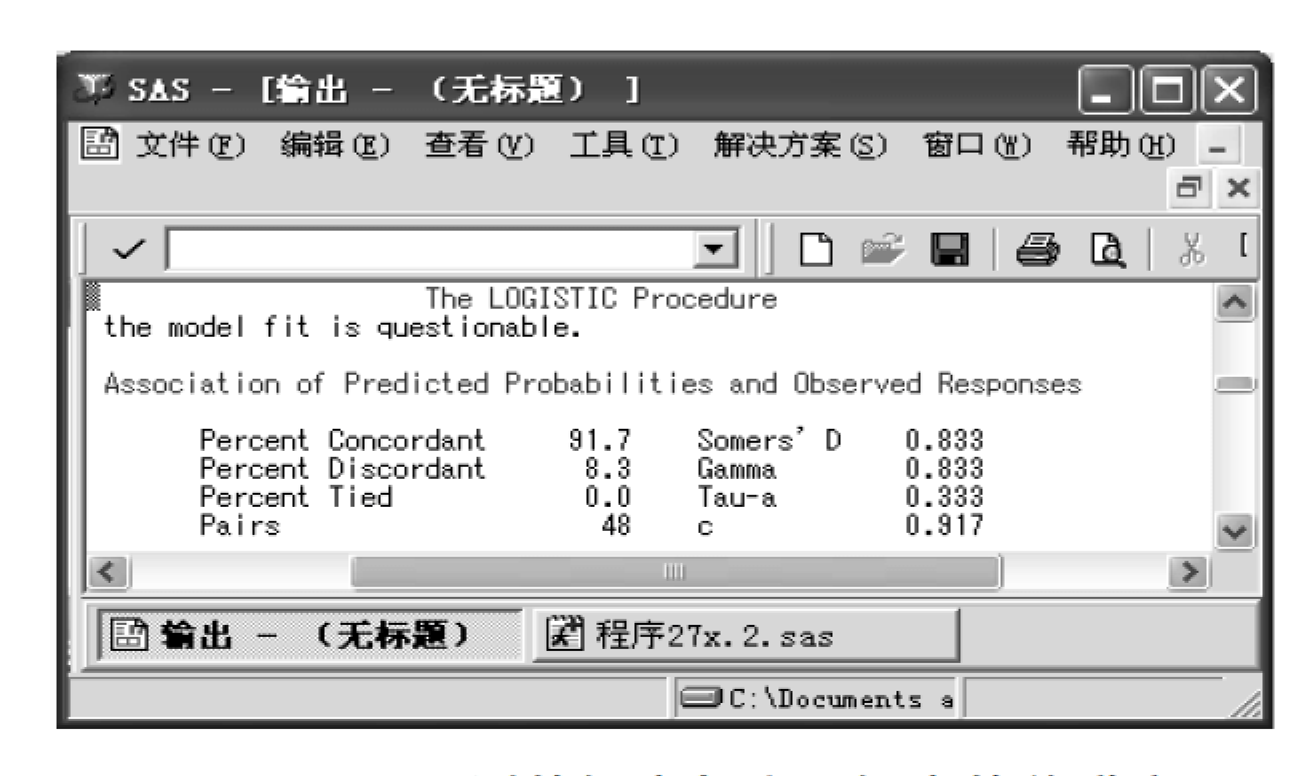
图 16.16　最大似然度的估计值



Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
AGE	94.630	<0.001	>999.999
LOCATION	0.578	0.006	57.180
SCORE	0.771	0.354	1.682
SEX	<0.001	<0.001	>999.999
HEIGHT	0.814	0.459	1.442

图 16.17　优势率的估计值



The LOGISTIC Procedure

the model fit is questionable.

Association of Predicted Probabilities and Observed Responses

Percent Concordant	91.7	Somers' D	0.833
Percent Discordant	8.3	Gamma	0.833
Percent Tied	0.0	Tau-a	0.333
Pairs	48	c	0.917

图 16.18　预测的概率与实际概率的关联度

这些图形的分析见 16.5 节。

16.5 假设与检验

1. 回归系数的假设检验

(1) 对于规模较大的样本(如 300 个 OBS 或更多 OBS),可利用 Wald 统计量检验一个回归系数是否为 0。

$$\text{Wald}=(\text{B}/(\text{S. E}))^2$$

(2) 回归系数的假设。

H_0(原假设):回归系数全部为 0;

H_1(备择假设):回归系数不是全部为 0。

(3) 回归系数的检验。

如图 16.16 所示,以截距项为例,回归系数 B=-25.6076,它的标准误差 S. E=184.7。因此:

$$\text{Wald}=(-25.6076\div 184.7)^2=0.0192 \tag{16.7}$$

Wald 的显著性水平见 Pr>ChiSq(相当于显著性水平 Sig)。在这个例子中,所有变量的 Wald 的显著性水平都大于 α 值 0.05,所以没有足够的理由拒绝“回归系数为 0”的原假设。说明各个变量的回归系数不显著,严格说来模型欠佳。

2. 为何不能拒绝原假设

当回归系数的绝对值变大时,其标准误差必然太大,因此使得 Wald 值变得过于小,以致于不能拒绝“系数为 0”的原假设。因此,无论什么时候,一旦回归系数变大,则不能依靠 Wald 值来检验假设,而应该根据图 16.15 中的 $-2\log L$(L 是 Likelihood 的简写)值的变化来检验假设。图 16.8 和图 16.16 中的回归系数尚未变大。

3. 偏相关系数 *R*

与其他多元回归一样,在 Logistic 回归中,对于一个单独的变量,很难判断它对模型的贡献。因为每个变量对模型的贡献率,还要依赖于模型中的其他变量。特别是在自变量之间有高度相关的情形下,更难计算出单个变量的贡献率。

因此,可以用统计量 R,来衡量每一个自变量对于因变量之间的偏相关系数。

R 的范围在(-1)~(+1)之间。当 $R>0$ 时,表示变量对模型的贡献率。但 R 越小,该变量对模型的贡献也就越小。

$$R=[(\text{Wald 值}-2K)/(-2\log L)]^{1/2} \tag{16.8}$$

式(16.8)中,K 是变量的自由度;$-2\log L$ 是 log likehihood 的缩写,而且 $-2\log L$ 值等于上一个模型的“log L 值”减去当前模型(包括截距+协变量)的“ log L 值”的差。

式(16.8)中的“$2K$”,是一个修正值,当 Wald 小于或等于 $2K$ 时,$R=0$。但在本例的输出结果中未显示 R 值。

16.6 解释回归系数

在多元 Logistic 回归中,对回归系数的解释很简单,它表示由于自变量的变化而引起因变量变化了多少。

1. 用 Log 值解释 Odds 值(优势值)

Odds 值(优势值)表示事件发生的优势值。为了便于进一步解释 Logistic 回归系数,下面改用 Odds 值取对数值来改写模型。所谓事件发生的 Odds 对数值,等于事件发生的概率除以事件未发生的概率,再对"商"求对数,即

$$
\begin{aligned}
(\text{Odds}) &= \log[\text{Prob(event)} \div \text{Prob(no event)}] \\
&= B_0 + B_1X_1 + B_2X_2 + \cdots + BpXp \qquad (16.9)
\end{aligned}
$$

式(16.9)中的 Logistic 系数可解释为:自变量的变化引起了 log 值的变化。如图 16.16 所示,Location 变量的系数为−0.5474,它表明地区变量从 0 变到 1 时,变量 V(恋爱)的 log 值增长了"−0.5474"。但是这种比较还不太直观,需要进一步用"概率比"解释回归模型。

2. 用"概率比"解释回归模型

为了便于进一步解释 Logistic 回归系数,还可将式(16.9)改写为事件发生与未发生的比例关系:

$$
\begin{aligned}
\text{Prob(event)}/\ \text{Prob(no event)} &= e^{B_0+B_1X_1+B_2X_2+\cdots+BpXp} \\
&= e^{B_0} * e^{B_1X_1} * e^{B_2X_2} * \cdots * e^{BpXp} \qquad (16.10)
\end{aligned}
$$

式(16.10)中,e^{B_i} 表明当第 i 个自变量值从 0 变化到 1 时,Odds 值变化了 e^{B_i} 倍。

如图 16.16 所示:对于 Location(地区)二分变量,当变量水平值从低(例如 0=农村)变到高(例如 1=城市)时,恋爱发生的概率反而递减(系数为−0.5474)。

16.7 发掘概率

在回归模型拟合数据的前提下,根据式(16.4)和图 16.16 中的回归系数,可以挖掘出如下恋爱发生的概率模型:

$$
\begin{aligned}
Z &= B_0 + B_1X_1 + B_2X_2 + \cdots + BpXp \\
&= -25.6076 + 4.55 * \text{age} - 0.5474 * \text{location} - 15.0249 * \text{sex} \\
&\quad - 0.2061 * \text{height} - 0.2596 * \text{score} \qquad (16.11)
\end{aligned}
$$

例 1:计算 22 岁农村男生、身高 170cm、各科平均成绩 80 分的恋爱概率。

解:

$$
\begin{aligned}
Z &= -25.6076 + 4.55 * \text{age} - 0.5474 * \text{location} - 15.0249 * \text{sex} \\
&\quad - 0.2061 * \text{height} - 0.2596 * \text{score}
\end{aligned}
$$

$$=-25.6076+4.55*22-0.5474*0-15.0249*1$$
$$-0.2061*170-0.2596*80$$
$$=-25.6076+100.1-0-15.0249-35.037-207.68$$
$$=-157.6419$$

把 $Z=-157.6419$ 代入式(16.4)得：

该生恋爱概率＝Prob(event)＝$1\div(1+e^{-z})=1\div(1+e^{157.6419})\approx 0$

当 Prob(event)≥0.5 时，事件将发生；当 Prob(event)＜0.5 时，事件不发生。所以该生可能未恋爱。

例 2：计算 22 岁城市男生、身高 170cm、各科平均成绩 80 分的恋爱概率。

解：

$$Z=-25.6076+4.55*age-0.5474*location-15.0249*sex$$
$$-0.2061*height-0.2596*score$$
$$=-25.6076+4.55*22-0.5474*1-15.0249*1$$
$$-0.2061*170-0.2596*80$$
$$=-25.6076+100.1-0.5474-15.0249-35.037-207.68$$
$$=-183.7969$$

把 $Z=-183.7969$ 代入式(16.4)得：

该生恋爱概率＝ Prob(event)＝ $1\div(1+e^{-z})=1\div(1+e^{183.7969})\approx 0$

当 Prob(event)≥0.5 时，事件将发生；当 Prob(event)＜0.5 时，事件不发生。所以该生可能未恋爱。

思考题：请计算 18 岁农村男生、身高 170cm、各科平均成绩 80 分的恋爱概率。

16.8 多分变量的编码

所谓二分变量，是将因变量的值编码成 1(男)和 0(女)两种值，在许多场合，是用“1”表示事件已发生，用“0”表示事件尚未发生。例如，在医学上用“1”表示淋巴结癌已扩散，用“0”表示淋巴结癌尚未扩散；或用“1”表示妊娠实验为阳性反应，用“0”表示妊娠试验呈阴性反应等。

但在某些场合，因变量有 3 种值以上，即多分变量(Categorical Variable)，这时要把因变量处理为哑变量(Dummy)或指示变量(Indicator)。请看下面两种编码设计。

1. 指示变量(Indicator-Variable)的编码方案

有 n 个数值的标称变量重新变换成($n-1$)个新变量时，则($n-1$)个新变量就成为指示变量。例如，在淋巴结癌细胞扩散试验中，当进行血清硫酸盐化学实验时，出现 3 种化验值，即 acid＝1(低值)，acid＝2(中值)，acid＝3(高值)。

但因 Logistic 回归有条件限制，当某变量的值多于 2 个时，就必须将它变为 $n-1$ 个指示变量。因此对 acid 变量的这 3 个值，必须改编为指示变量 CATacid1、CATacid2 等。

当新变量 CATacid1＝1 时是低值，CATacid1＝0 时不是低值；当新变量 CATacid2＝

1 时是中值,CATacid2=0 时不是中值。

原始变量 ACID 的高值,虽然没有专门建立新变量,但由 CATacid1 和 CATacid2 这两个新变量同时为 0 编码时表示“化验值”为高值,因为既不是中值又不是低值,必然是高值,即当新变量 CATacid1=0,同时新变量 CATacid2=0 时,则隐含 acid=3(高值),见图 16.19(a)。

Value	CATacid	
	(1)	(2)
低值	1.000	0.000
中值	0.000	1.000
高值	0.000	0.000

(a) 对旧变量的改编

	Value	Freq	Parameter coding	
			(1)	(2)
CATacid				
	1.00	25	1.000	0.000
	2.00	26	0.000	1.000
	3.00	18	−1.000	−1.000

(b) 对旧变量的另一种编码方案

图 16.19 多分变量的编码

2. 另一个编码方案

图 16.19(a)是用指示变量 CATacid1 的“1”表示旧变量 acid 的低值,用 CATacid2 的“1” 代表旧变量 acid 的中值,当这 2 个指示变量的代码同时为“0”时表示旧变量 acid 的高值。

下面改用另一种编码方案,其思路是将血清反应的旧变量 acid 的每一个值的系数,与总平均系数作比较,其编码方案见图 16.19(b)。

如图 16.19(b)所示,旧变量(acid)原来的高值(代码 3.00),现在被重新编码为−1.000和−1.000 两个值,而不是 0.000 和 0.000。也就是说,此时是将这 2 个指示变量 CATacid1 和 CATacid2 中的“−1”代码作为旧变量 acid 的高值。

用图 16.19(b)的编码方案后,指示变量的某一个数值所对应的结果,将与总平均结果进行比较。

习 题 16

1. 多个自变量的 Logistic Regression 模型是什么?
2. 二分的 Logistic Regression 回归对因变量和自变量各有什么要求?
3. Logistic Regression 输出的回归系数和截距是什么值?

4. 下面是 Logistic Regression 回归的常识问题。

(1) 预测一个事件是否已经发生，以及判别“一个因变量对于这种预测”的结果如何等，为什么要用 Logistic 回归法?

(2) 对于规模较大的样本，可利用什么统计量检验一个回归系数是否为 0?

(3) 当回归系数的绝对值变大时，为什么不能拒绝“系数为 0”的原假设?

(4) 当回归系数的绝对值变大时怎么办?

(5) 如何才能更直接地解释回归系数?

5. 试举出一个用 PROC Logistic 编程法做 Logistic 回归的例子。

提示：参阅第 16 章 16.4 节的程序 16.2。

第 17 章 chapter 17

2 * 2 维 Logistic Regression 回归分析

第 16 章是探讨因变量只有二水平(0 为事件不发生、1 为事件已发生)、自变量为区间(定距)以上类型的 Logistic Regression 回归分析。但是,当因变量只有两个水平(0 为事件不发生、1 为事件已发生),而自变量也限制为两个水平(例如:0 为疾病不知道或未暴露,1 为疾病已知道或已暴露)时,则成为 2×2 维变量,即 2 * 2 维变量。例如医学上,因变量为某种疾病(0 为不发生、1 为发生),自变量为吸烟状况(0 为不吸烟、1 为吸烟)时是以成对的形式出现的。

又例如,社会上青少年犯罪与父母是否离异,二者之间也是以成对的数据出现的。

类似这样的数据被称为"比较研究"数据,这类数据可以用 Logistic Regression 过程加以分析,也就是常说的 2 * 2 维(或方表)Logistic Regression 回归分析。

17.1 2 * 2 维 Logistic Regression 模型

2 * 2 维 Logistic Regression 模型,和第 16 章通用的 Logistic Regression 回归模型完全一样,见式(17.1)。

$$\text{Prob(event)} = 1/[\ 1 + e^{-(B_0+B_1\times X_1)}\ ] \tag{17.1}$$

式(17.1)中:

B_0:回归截距。

B_1:从数据中计算出的回归系数。

X_1:自变量。

e:自然对数的底,e≈2.178。

17.2 2 * 2 维 Logistic Regression 的变量及其数据

1. 变量和数据:见表 17.1。

表 17.1 变量定义

犯 罪 情 况	看录像(编码 1)	不看录像(编码 0)
犯罪(编码 1)	25	15
不犯罪(编码 0)	50	40

2. 在程序编辑器窗口输入数据:见程序 17.1。

程序 17.1:

```
DATA lux;
INPUT fz lx freq;
CARDS;
1 1 25
1 0 15
0 1 50
0 0 40
;
RUN;
```

请上机运行程序 17.1,然后将输出结果与 17.3 节的对话框中所产生的结果进行对比。

下面 17.3 节是改用选择对话框的命令来运行程序 17.1 中的数据。

17.3 用"分析家"对话框进行 2 * 2维 Logistic 回归

1. "分析家"对话框的操作步骤

(1) 选择图 17.1 的 SAS 主菜单中的"运行"→"提交"命令运行程序 17.1 及其数据,产生 SAS 数据集 Work. lux。

(2) 选择图 17.1 所示的"解决方案"→"分析"命令,鼠标指针移到图 17.1 中带有阴

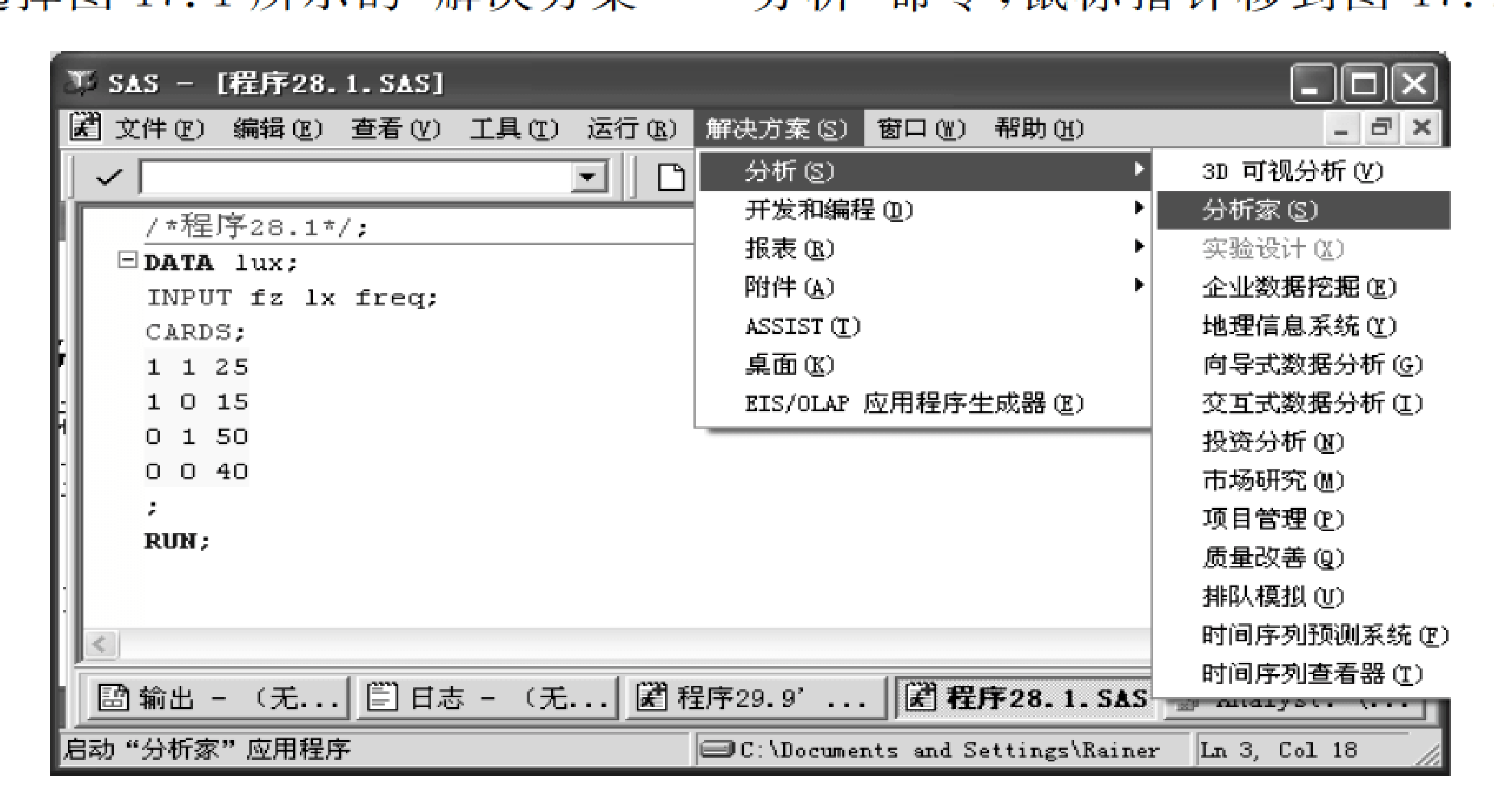


图 17.1 分析家的菜单位置

影标记的“分析家”命令上。

（3）单击图 17.1 中的“分析家”→“文件”→“按 SAS 名称打开”→Work→lux 命令按钮，进入图 17.2 所示的对话框。



图 17.2 Work.lux 数据集文件

（4）选择文件名 lux 后单击“确定”按钮，出现图 17.3 所示的 Work.lux 数据集数据。



图 17.3 Work.lux 的文件内容

（5）选择“统计”→“回归”命令进入图 17.4。

（6）选择 Logistic 命令后进入图 17.5 并设置变量。

说明： lx 是标称(定类)型变量只能放在 Class 文本框内。非标称变量必须放在图 17.5 的 Quantitative 框内。

（7）单击图 17.5 中的 Model 按钮，进入图 17.6 后单击 Standard Models 按钮，选择默认的主效应模型。



图 17.4　Logistic 过程的菜单位置

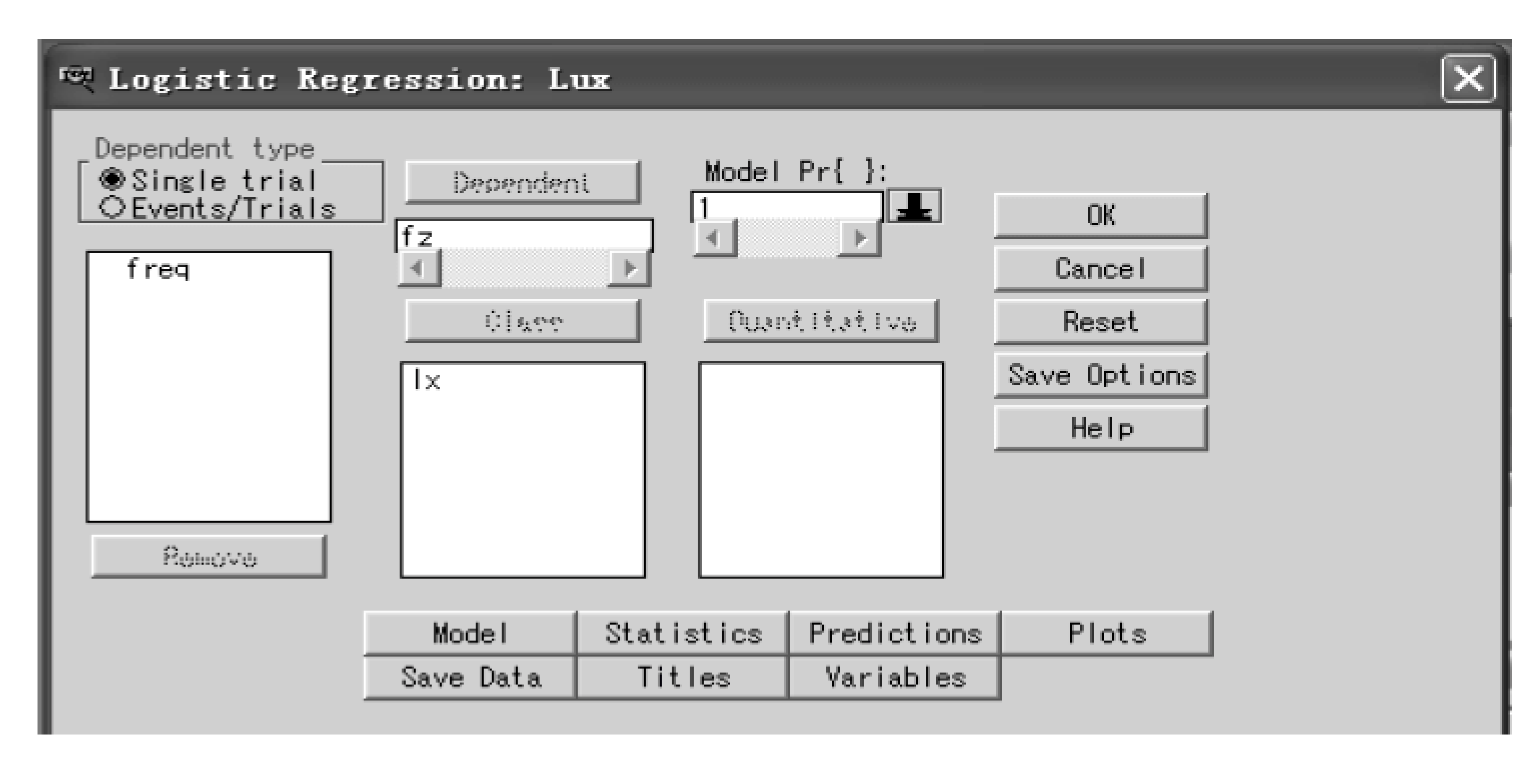


图 17.5　选择 Logistic 的因变量 fz(犯罪)和自变量 lx(录像)

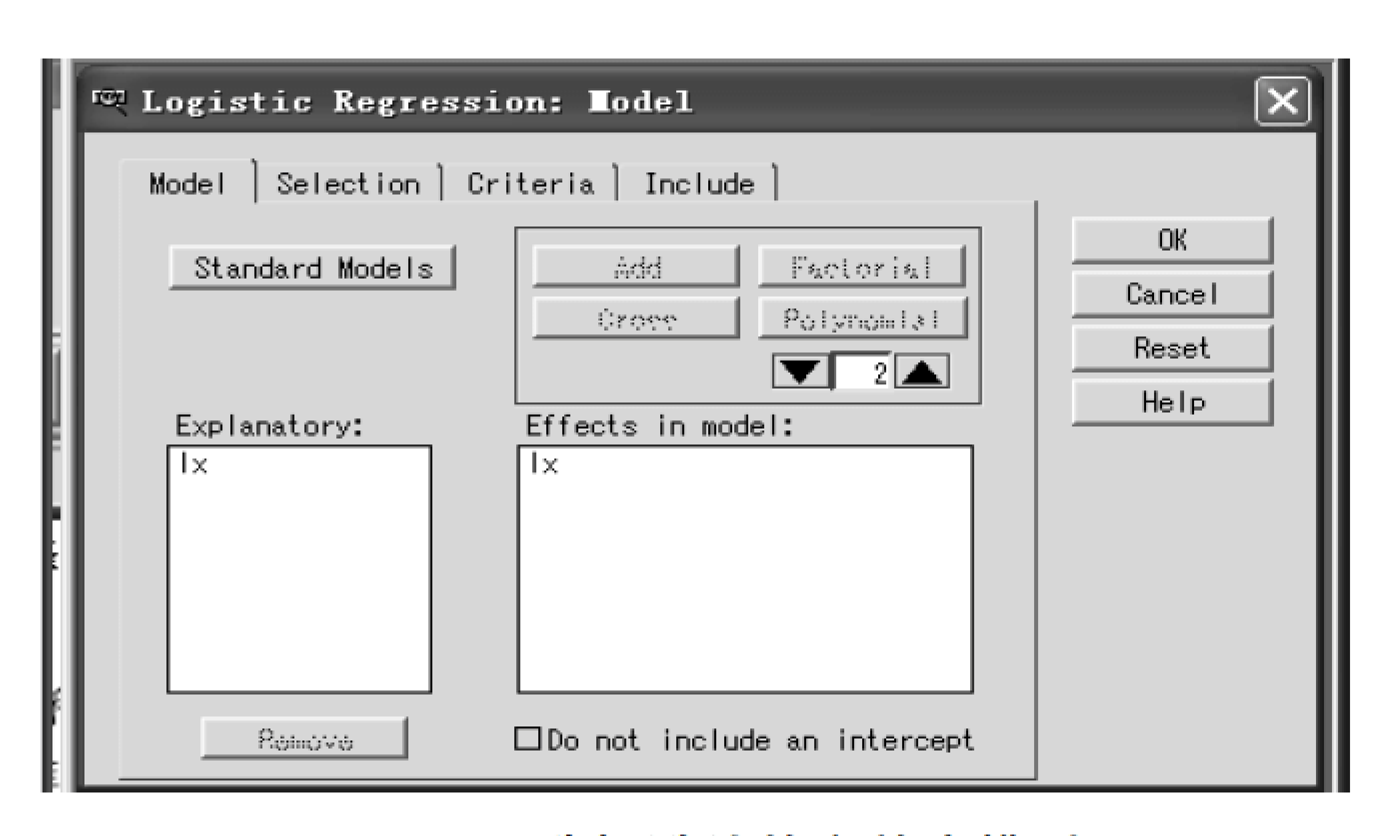


图 17.6　选择默认的主效应模型

(8) 单击 OK 按钮后再单击 Statistics 标签，进入图 17.7 选择默认的统计量。

(9) 图 17.7 的统计量可以不选(即按默认的)。单击 OK 按钮回到图 17.5。

(10) 单击图 17.5 下方的 Variable 按钮，进入图 17.8 后把变量 Freq 选为加权变量。此项必选。

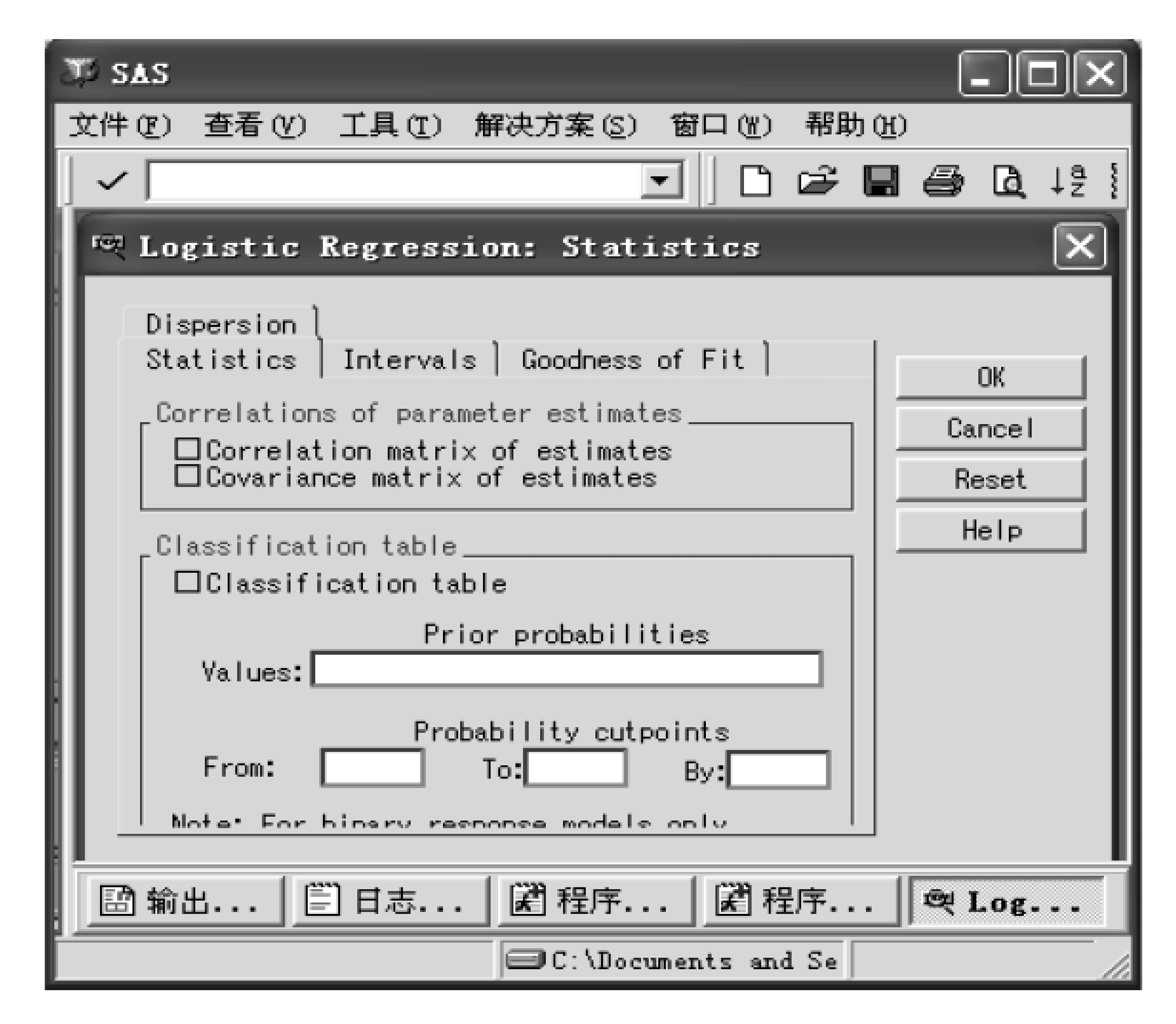


图 17.7 统计量对话框

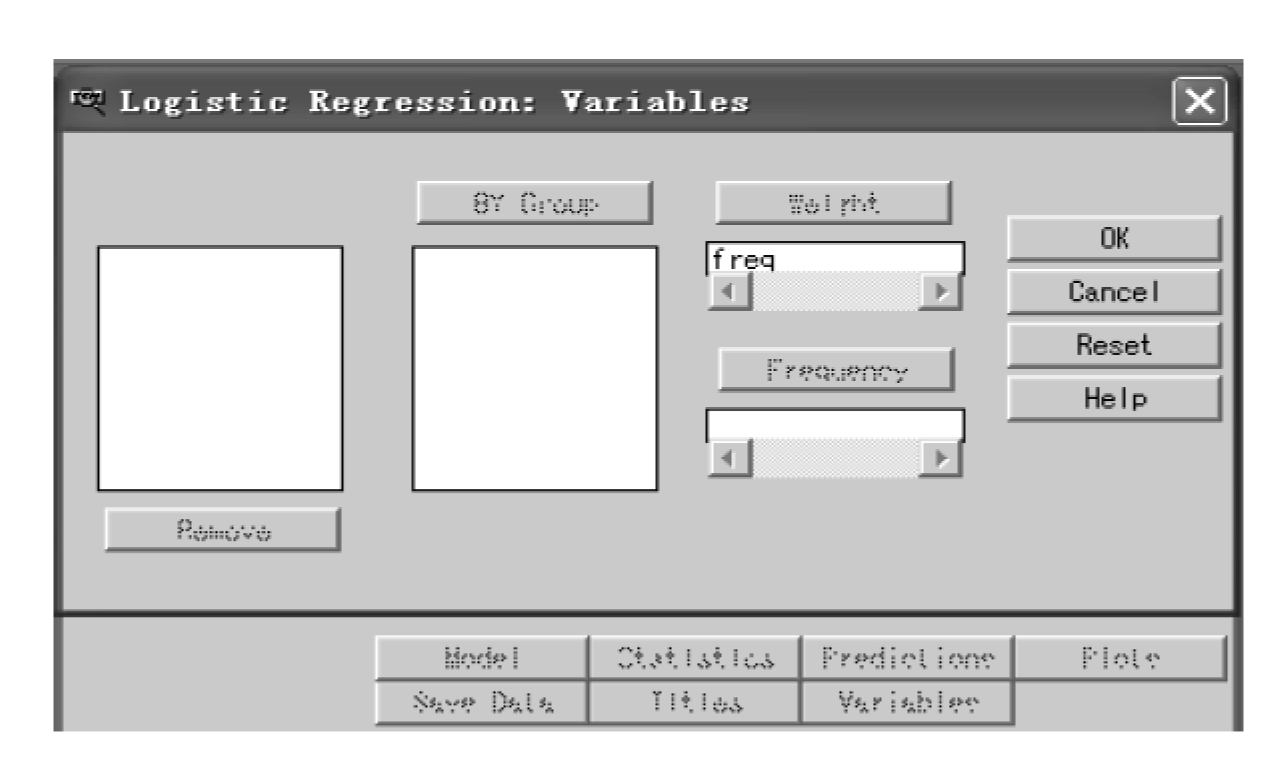


图 17.8 把变量 freq 选为"加权变量"

(11) 图 17.5 下方的其他对话框选项可以不选,单击两次 OK 按钮后产生图 17.9 至图 17.13 的输出结果。

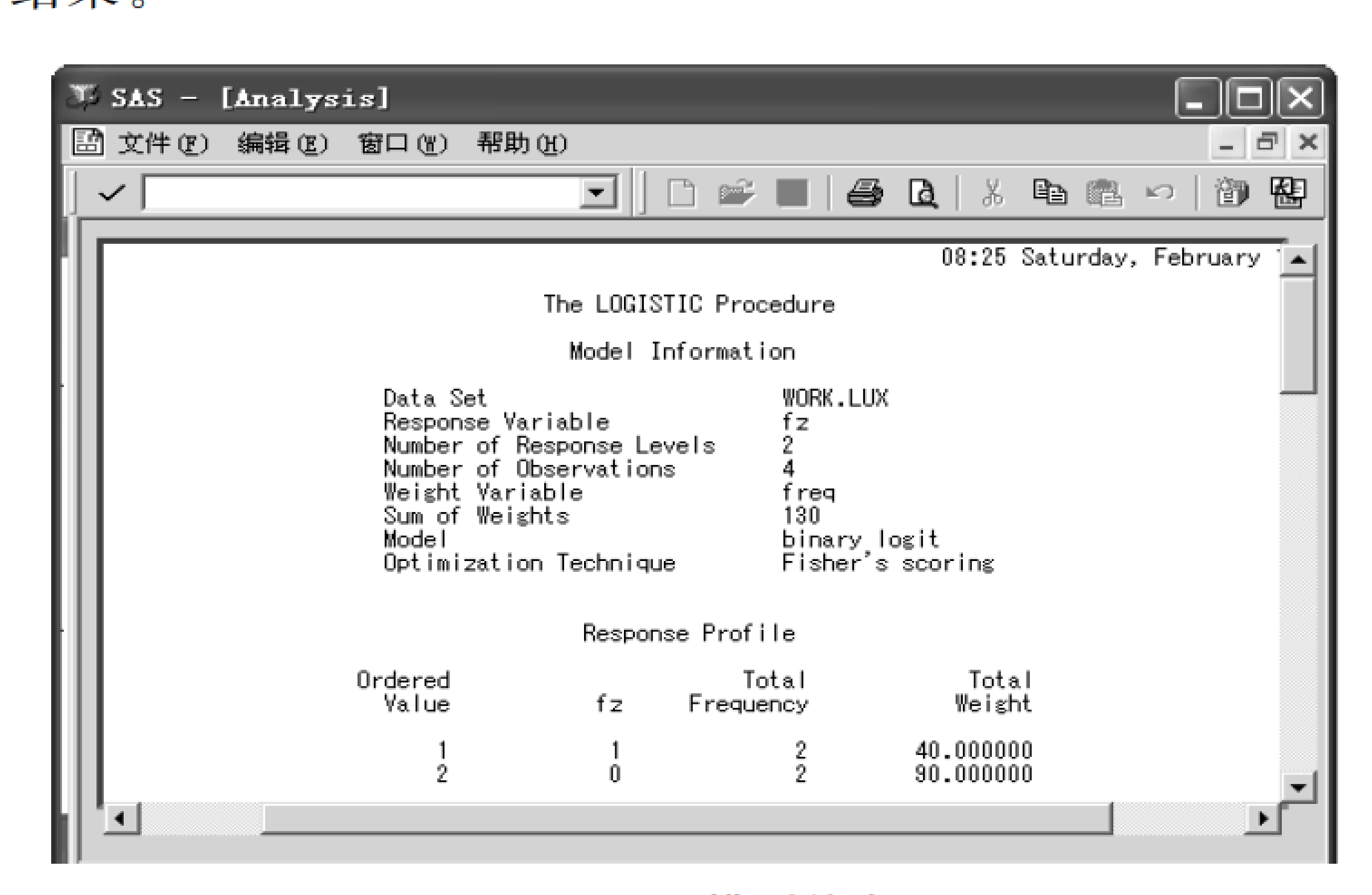
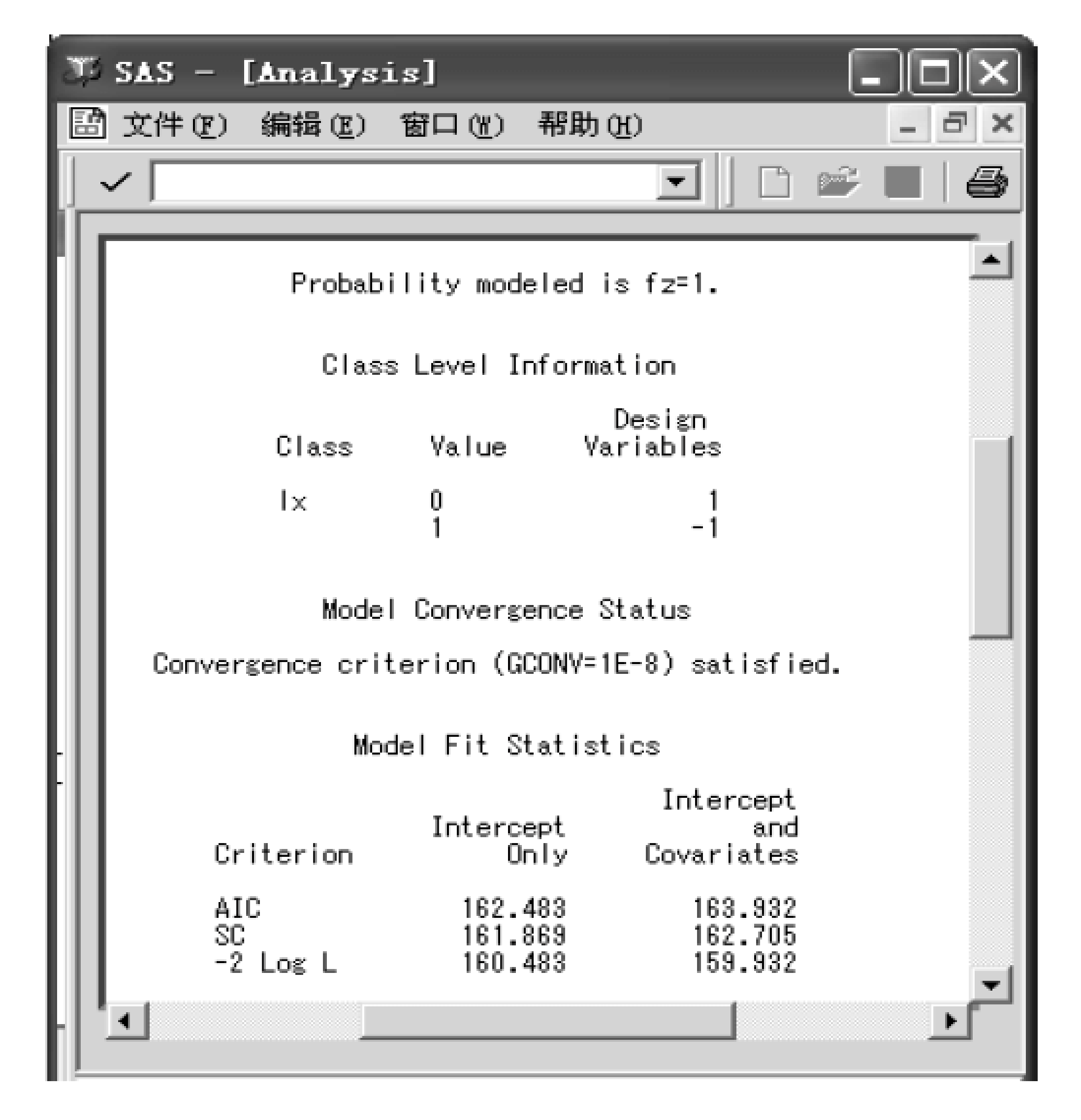


图 17.9 模型信息



SAS – [Analysis]

Probability modeled is fz=1.

Class Level Information

Class	Value	Design Variables
lx	0	1
	1	-1

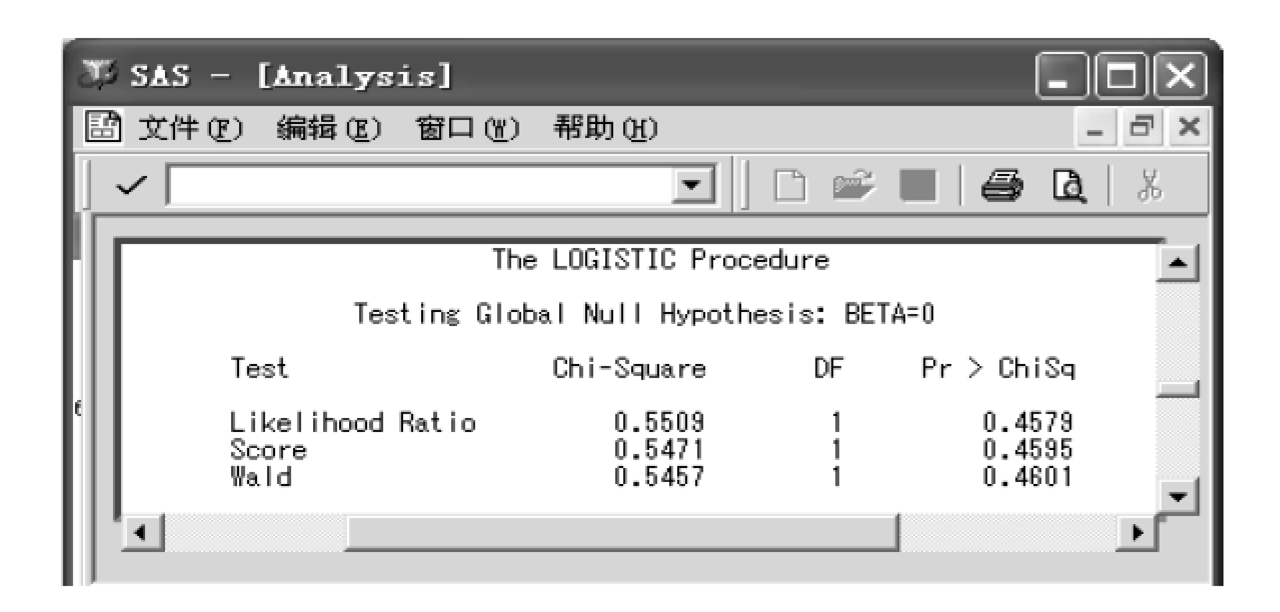
Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	162.483	163.932
SC	161.869	162.705
-2 Log L	160.483	159.932

图 17.10　模型拟合检验



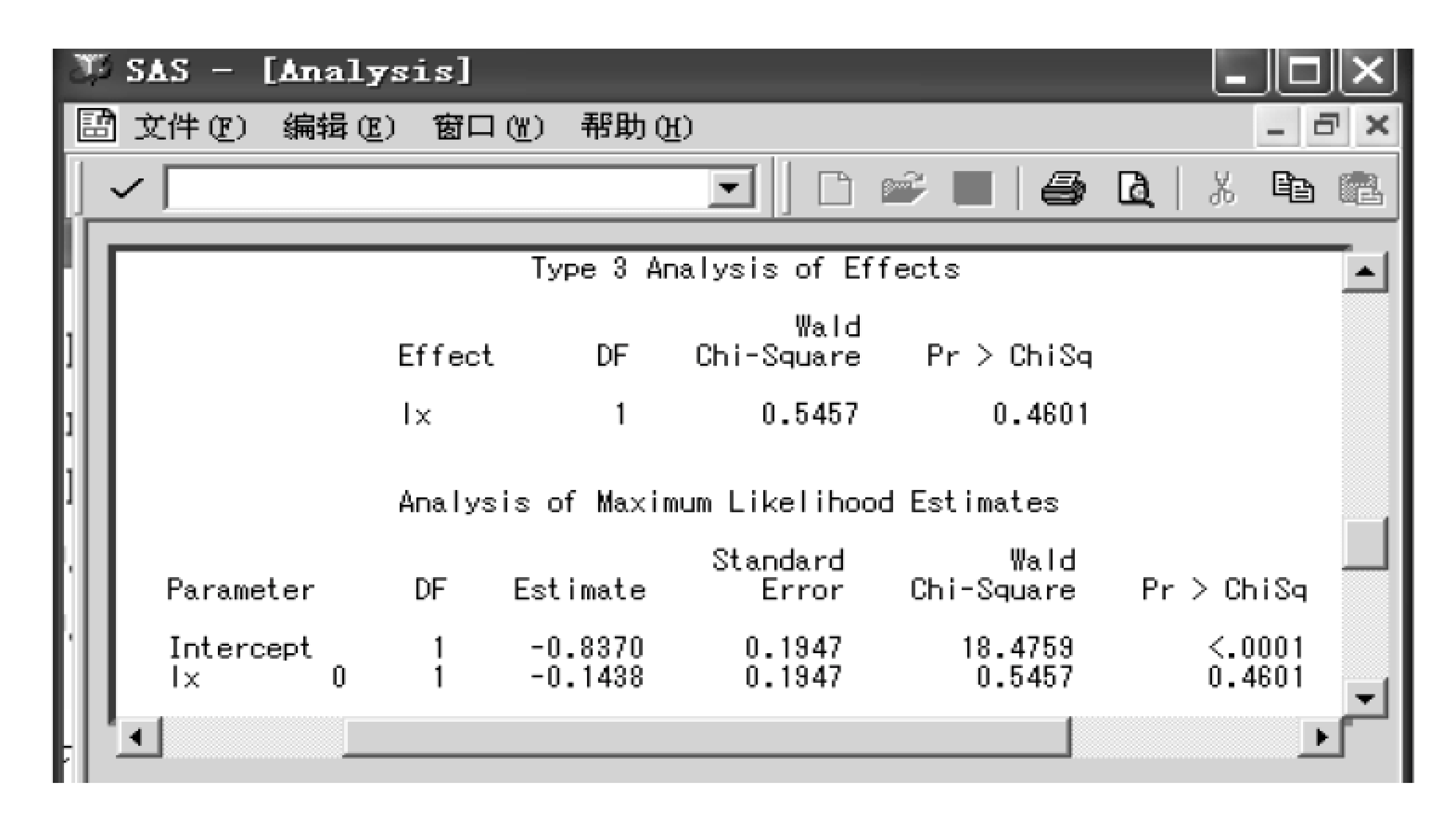
SAS – [Analysis]

The LOGISTIC Procedure

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	0.5509	1	0.4579
Score	0.5471	1	0.4595
Wald	0.5457	1	0.4601

图 17.11　系数全 0 的假设检验



SAS – [Analysis]

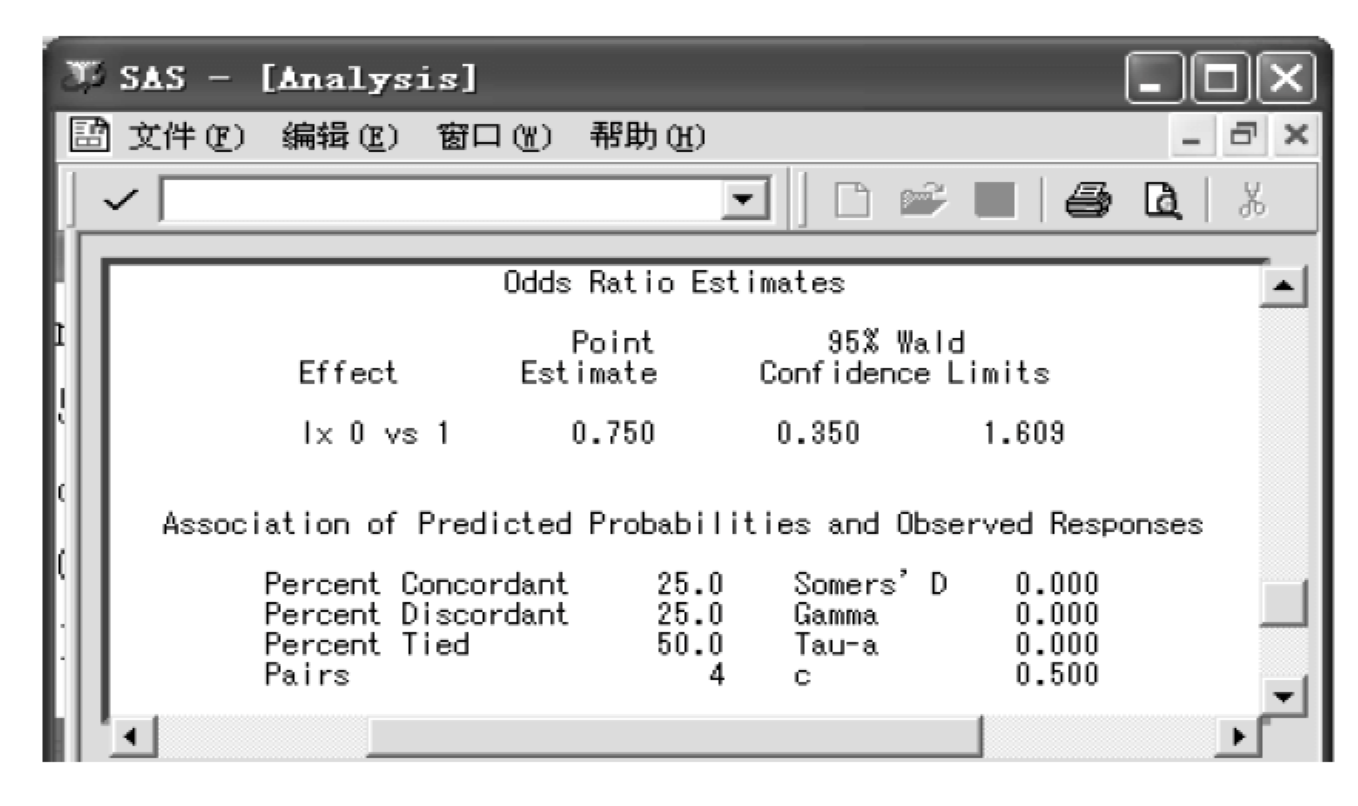
Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
lx	1	0.5457	0.4601

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-0.8370	0.1947	18.4759	<.0001
lx	0	1	-0.1438	0.1947	0.5457	0.4601

图 17.12　最大似然率估计



SAS – [Analysis]

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
lx 0 vs 1	0.750	0.350	1.609

Association of Predicted Probabilities and Observed Responses

Percent Concordant	25.0	Somers' D	0.000
Percent Discordant	25.0	Gamma	0.000
Percent Tied	50.0	Tau-a	0.000
Pairs	4	c	0.500

图 17.13　dds(优势率)估计

2. 输出结果分析：见 17.4 节

17.4 2＊2维 Logistic 回归分析

1. 模型拟合度

从图 17.10 的－2 log L 值看：模型不拟合数据。

2. 回归系数为 0 的假设检验

从图 17.11 和图 17.12 看：变量 Lx 的“Pr＞ChiSq”值大于 α 值 0.05，所以没有理由拒绝“系数为 0 的假设”。

3. 回归模型

为了能继续向下进行讲解，特假定系数不为 0，于是，从图 17.12 可获得 Logistic 回归模型如下。

$$P = 1 \div (1 + e^{-(-0.8370-0.1438 \times Lx)}) = 1 \div (1 + e^{0.8370+0.1438 \times Lx}) \qquad (17.2)$$

4. 概率预测

根据图 17.10 的编码可进行以下的概率预测：

(1) 不看录像的犯罪率为 $P1 = 1 \div (1 + e^{0.837+0.1438 \times 1}) = 1 \div (1 + e^{0.9808}) = 1 \div (1 + 2.67) = 0.27$

(2) 看录像的犯罪率为 $P2 = 1 \div (1 + e^{0.837+0.1438 \times 0}) = 1 \div (1 + e^{0.837}) = 1 \div (1 + 2.31) = 0.30$

上述 0.27 与 0.30 这两种比例失调显然不能令人满意，其原因之一是数据本身缺乏随机性，但是所讲授的回归方法不失一般意义。

习 题 17

1. 什么是 2＊2 维 Logistic Regression 模型？

2. 表 17.2 是某单位 200 名职工中 20 年间抽烟与否和心脏病的关系数据，请建立 2＊2维 Logistic Regression 模型，并且计算 Odds 率。

表 17.2 抽烟与否和心脏病的关系数据

	抽烟(编码 chy ＝1)	**不抽烟(编码 chy ＝0)**
有心脏病 (编码 xzb ＝1)	68 人	32 人
无心脏病 (编码 xzb ＝0)	18 人	82 人

提示：2 * 2 维 Logistic Regression 模型见程序 17.2。

程序 17.2：

```
DATA XZB0;
INPUT xzb chy freq @ @ ;
CARDS;
1 1 68
1 0 32
0 1 18
0 0 82
;
PROC LOGISTIC DATA= xzb0;
MODEL xzb= chy;
WEIGHT freq;
RUN;
```

编码说明：xzb＝1(有心脏病)　xzb＝0(无心脏病)
　　　　　chy＝1(抽烟)　　chy＝0(不抽烟)

高等院校信息技术规划教材

系 列 书 目

书　名	书　号	作　者
数字电路逻辑设计	978-7-302-12235-7	朱正伟　等
计算机网络基础	978-7-302-12236-4	符彦惟　等
微机接口与应用	978-7-302-12234-0	王正洪　等
XML 应用教程(第 2 版)	978-7-302-14886-9	吴　洁
算法与数据结构	978-7-302-11865-7	宁正元　等
算法与数据结构习题精解和实验指导	978-7-302-14803-6	宁正元　等
工业组态软件实用技术	978-7-302-11500-7	龚运新　等
MATLAB 语言及其在电子信息工程中的应用	978-7-302-10347-9	王洪元
微型计算机组装与系统维护	978-7-302-09826-3	厉荣卫　等
嵌入式系统设计原理及应用	978-7-302-09638-2	符意德
C++ 语言程序设计	978-7-302-09636-8	袁启昌　等
计算机信息技术教程	978-7-302-09961-1	唐　全　等
计算机信息技术实验教程	978-7-302-12416-0	唐　全　等
Visual Basic 程序设计	978-7-302-13602-6	白康生　等
单片机 C 语言开发技术	978-7-302-13508-1	龚运新
ATMEL 新型 AT89S52 系列单片机及其应用	978-7-302-09460-8	孙育才
计算机信息技术基础	978-7-302-10761-3	沈孟涛
计算机信息技术基础实验	978-7-302-13889-1	沈孟涛　著
C 语言程序设计	978-7-302-11103-0	徐连信
C 语言程序设计习题解答与实验指导	978-7-302-11102-3	徐连信　等
计算机组成原理实用教程	978-7-302-13509-8	王万生
微机原理与汇编语言实用教程	978-7-302-13417-6	方立友
微机组装与维护用教程	978-7-302-13550-0	徐世宏
计算机网络技术及应用	978-7-302-14612-4	沈鑫剡　等
微型计算机原理与接口技术	978-7-302-14195-2	孙力娟　等
基于 MATLAB 的计算机图形与动画技术	978-7-302-14954-5	于万波
基于 MATLAB 的信号与系统实验指导	978-7-302-15251-4	甘俊英　等
信号与系统学习指导和习题解析	978-7-302-15191-3	甘俊英　等
计算机与网络安全实用技术	978-7-302-15174-6	杨云江　等
Visual Basic 程序设计学习和实验指导	978-7-302-15948-3	白康生　等
Photoshop 图像处理实用教程	978-7-302-15762-5	袁启昌　等
数据库与 SQL Server 2005 教程	978-7-302-15841-7	钱雪忠　著

读者意见反馈

亲爱的读者：

感谢您一直以来对清华版计算机教材的支持和爱护。为了今后为您提供更优秀的教材，请您抽出宝贵的时间来填写下面的意见反馈表，以便我们更好地对本教材做进一步改进。同时如果您在使用本教材的过程中遇到了什么问题，或者有什么好的建议，也请您来信告诉我们。

地址：北京市海淀区双清路学研大厦 A 座 602　　计算机与信息分社营销室 收
邮编：100084　　电子邮件：jsjjc@tup.tsinghua.edu.cn
电话：010-62770175-4608/4409　　邮购电话：010-62786544

教材名称：SAS 数据挖掘与分析
ISBN：978-7-302-16920-8

个人资料

姓名：＿＿＿＿＿＿＿　年龄：＿＿＿＿　所在院校/专业：＿＿＿＿＿＿＿＿
文化程度：＿＿＿＿＿＿　通信地址：＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿
联系电话：＿＿＿＿＿＿　电子信箱：＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿

您使用本书是作为：□指定教材 □选用教材 □辅导教材 □自学教材

您对本书封面设计的满意度：

□很满意 □满意 □一般 □不满意　改进建议＿＿＿＿＿＿＿＿＿＿＿＿＿＿

您对本书印刷质量的满意度：

□很满意 □满意 □一般 □不满意　改进建议＿＿＿＿＿＿＿＿＿＿＿＿＿＿

您对本书的总体满意度：

从语言质量角度看 □很满意 □满意 □一般 □不满意
从科技含量角度看 □很满意 □满意 □一般 □不满意

本书最令您满意的是：

□指导明确 □内容充实 □讲解详尽 □实例丰富

您认为本书在哪些地方应进行修改？（可附页）

＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿
＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿

您希望本书在哪些方面进行改进？（可附页）

＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿
＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿

电子教案支持

敬爱的教师：

为了配合本课程的教学需要，本教材配有配套的电子教案（素材），有需求的教师可以与我们联系，我们将向使用本教材进行教学的教师免费赠送电子教案（素材），希望有助于教学活动的开展。相关信息请拨打电话 010-62776969 或发送电子邮件至 jsjjc@tup.tsinghua.edu.cn 咨询，也可以到清华大学出版社主页（http://www.tup.com.cn 或 http://www.tup.tsinghua.edu.cn）上查询。